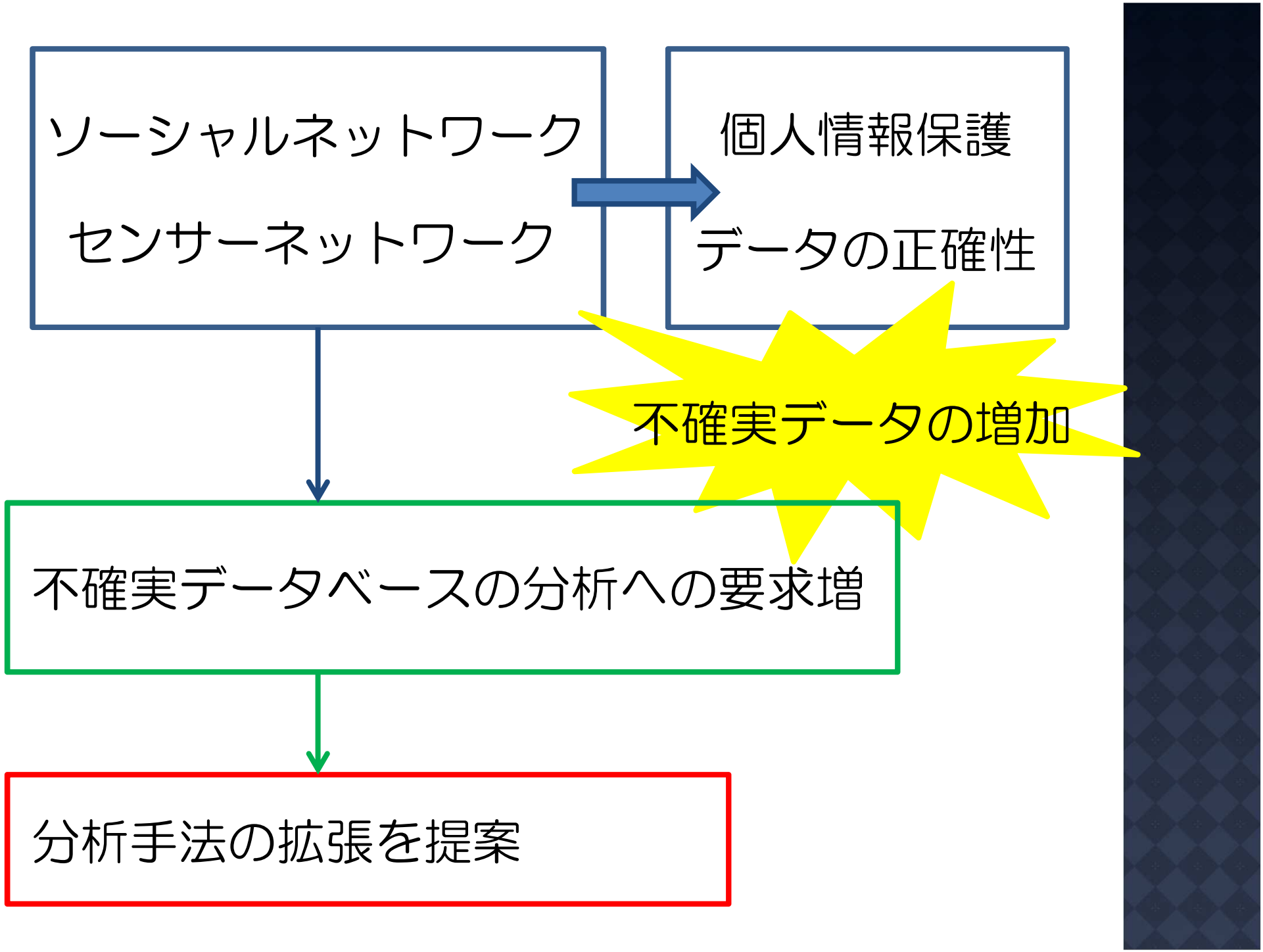
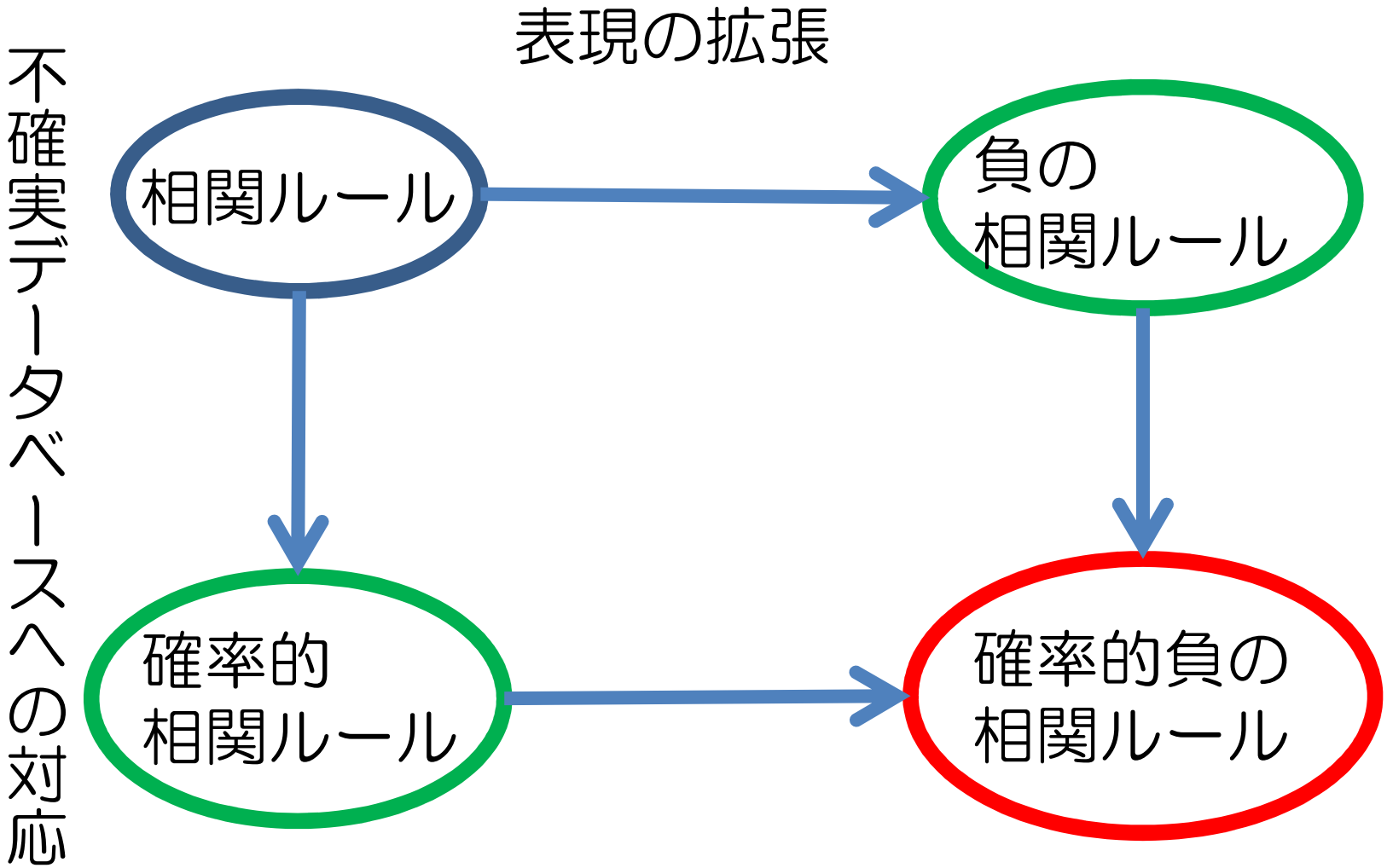
不確実データベースからの負の相関ルールの抽出

情報システム解析学科4年
藤田岳行


ソーシャルネットワーク
センサーネットワーク
個人情報保護
データの正確性
不確実データの増加
不確実データベースの分析への要求増
分析手法の拡張を提案

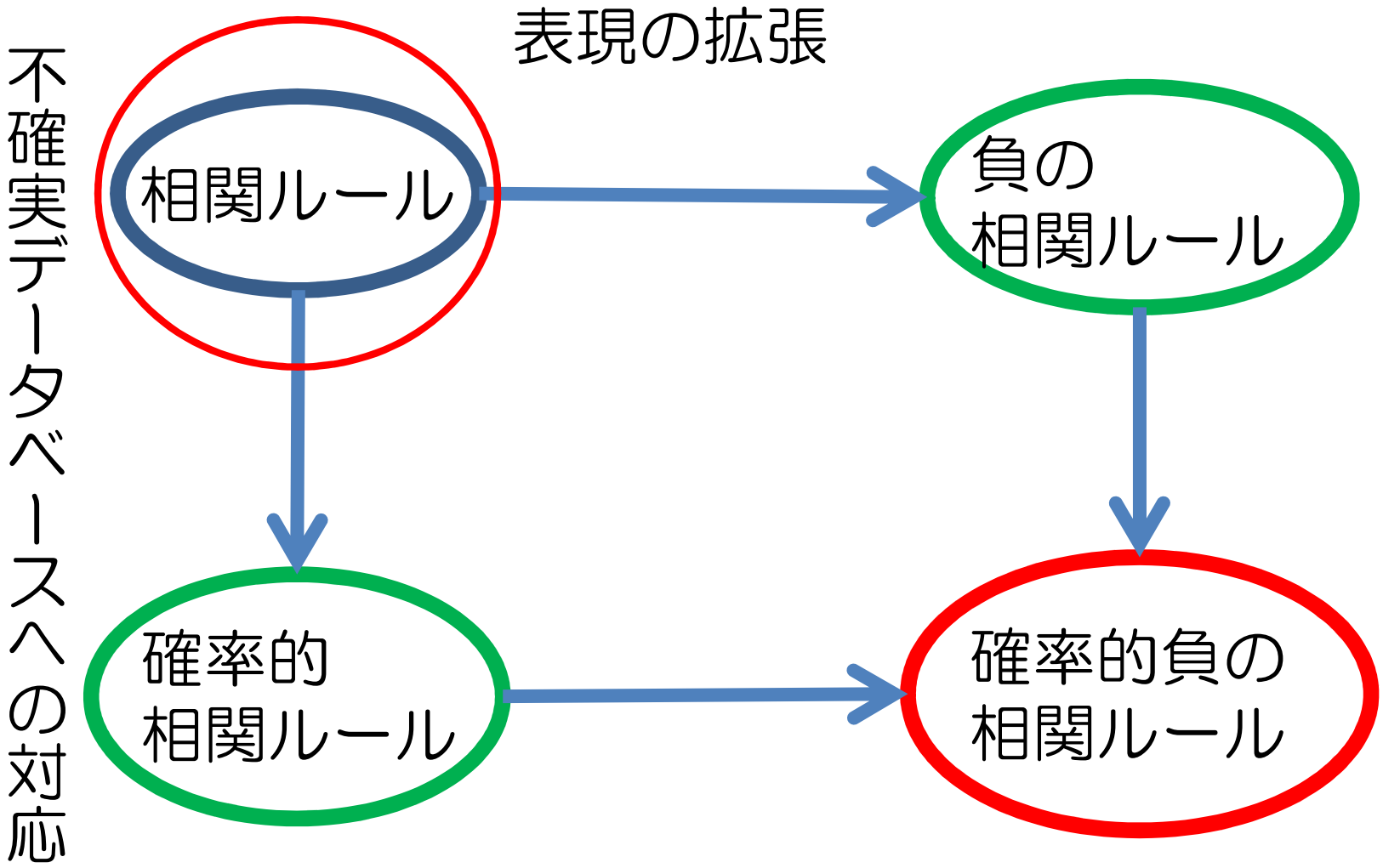
既存研究との関係



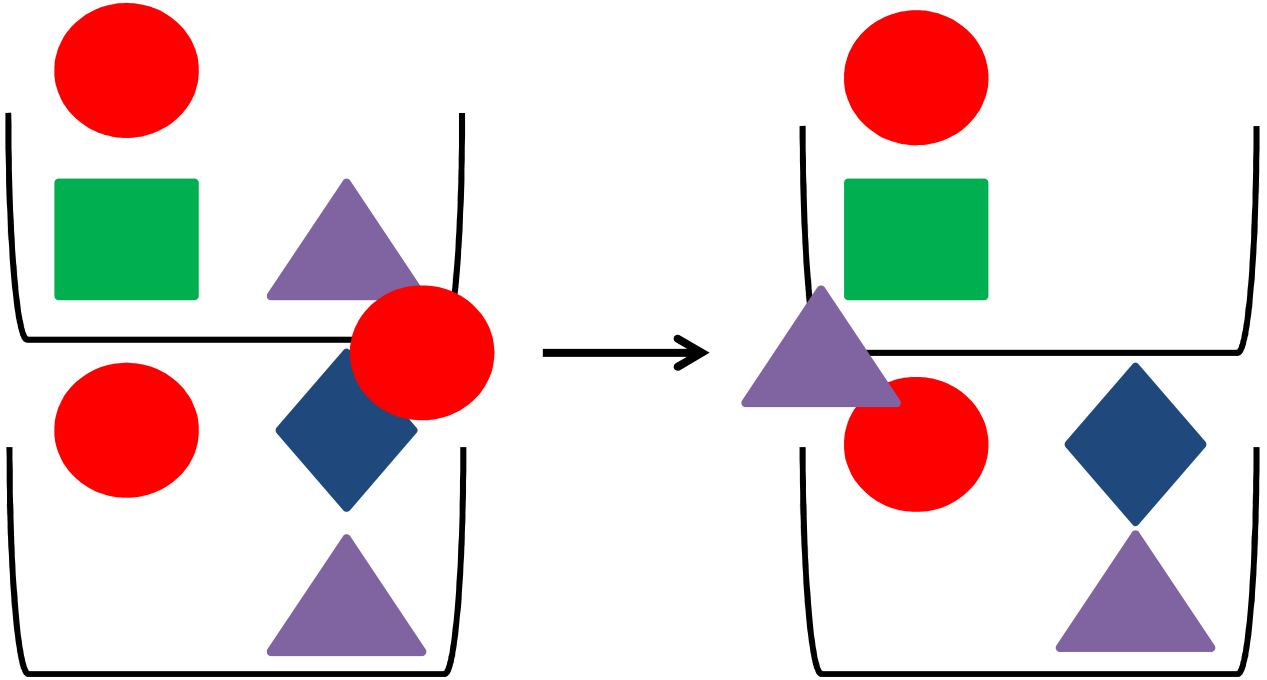
目次

- 相関ルール
- 負の相関ルール
- 確率的相関ルール
- 確率的負の相関ルール
  - 定義と計算方法
  - 探索の方針
- 実験と結果
- まとめ

既存研究


表現の拡張
不確実データベースへの対応
相関ルール
負の
相関ルール
確率的
相関ルール
確率的負の
相関ルール

相関ルール



2つの集合の共起をルールの形で表したもの

評価基準

支持度(同時確率)

$$sup(X \Rightarrow Y)$$

確信度(条件付き確率)

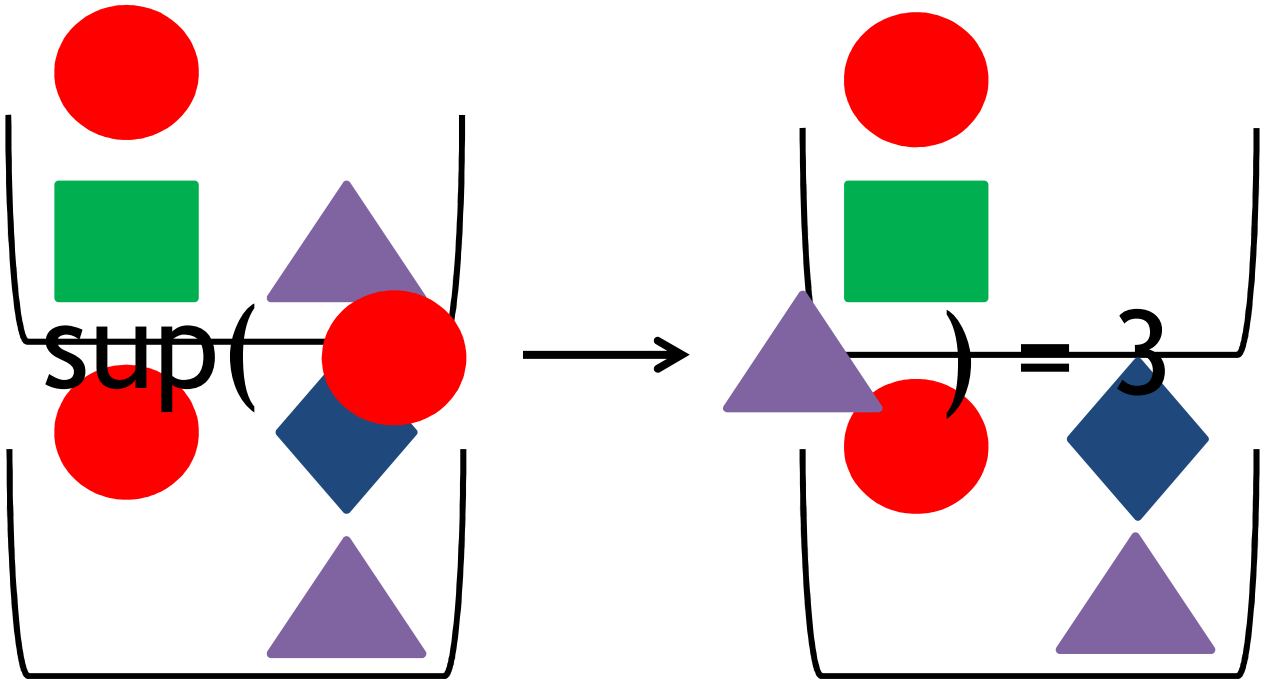
$$conf(X \Rightarrow Y)$$

前件

後件

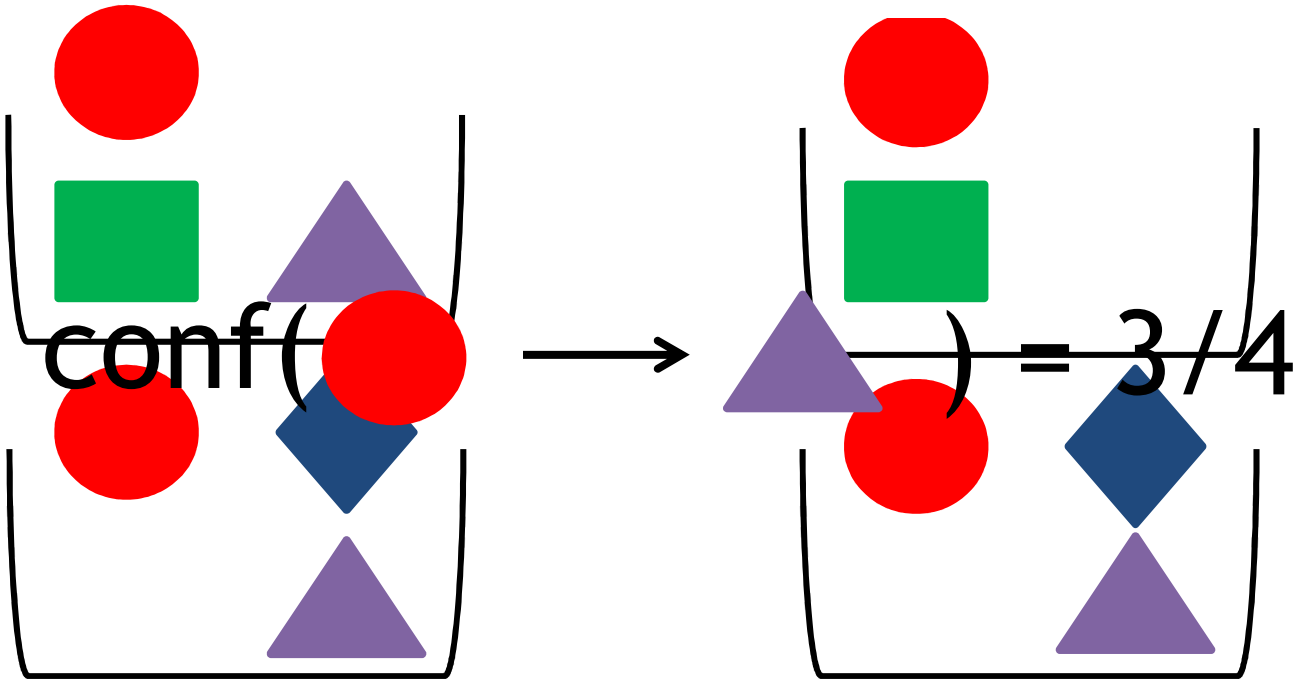
＊前件Xおよび後件Yは集合

$$sup(X \Rightarrow Y)$$



支持度とはパターンが出現した数である

$$conf(X \Rightarrow Y)$$

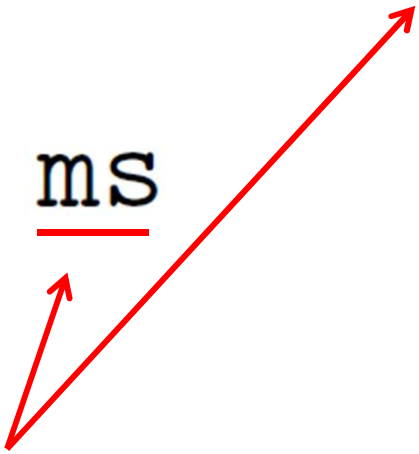


確信度とは前件が出現したうち後件が出現する確率である

有効な相関ルールの定義

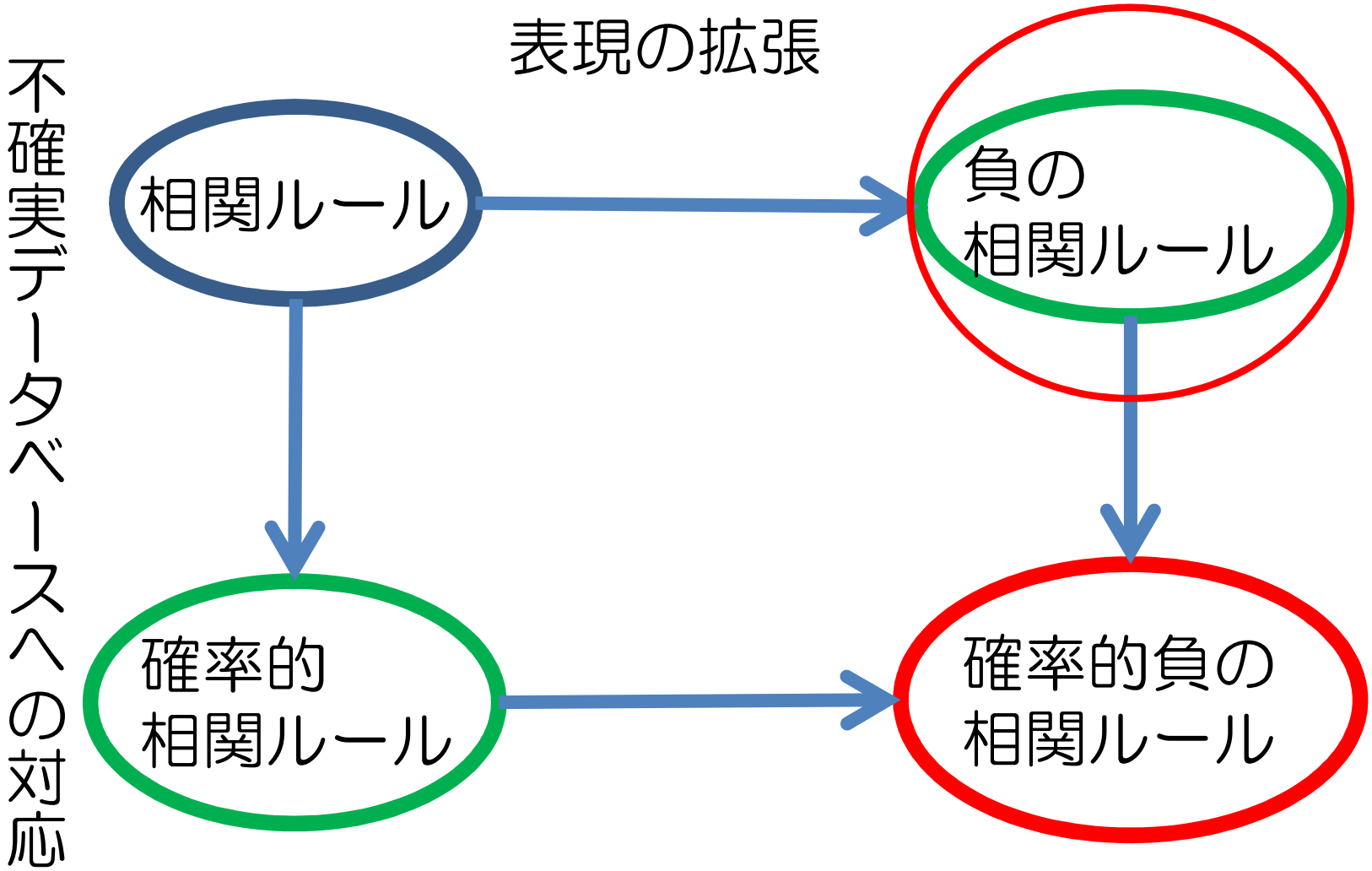
$$conf(\ X \Rightarrow Y\ ) \geq \mathtt{mc}$$

$$sup(XY) \geq \mathtt{ms}$$



mc(最小確信度),ms(最小支持度)は
ユーザーが任意に定める閾値

支持度においてX⇒YとXYは同じである


不確実データベースへの対応
表現の拡張
相関ルール
負の
相関ルール
確率的
相関ルール
確率的負の
相関ルール

負の相関ルール

出現(X)と否出現(¬X)の組み合わせをルールとして表したもの

$$X \Rightarrow \neg Y$$

$$\neg X \Rightarrow Y$$

$$\neg X \Rightarrow \neg Y$$

有効な負の相関ルールの定義

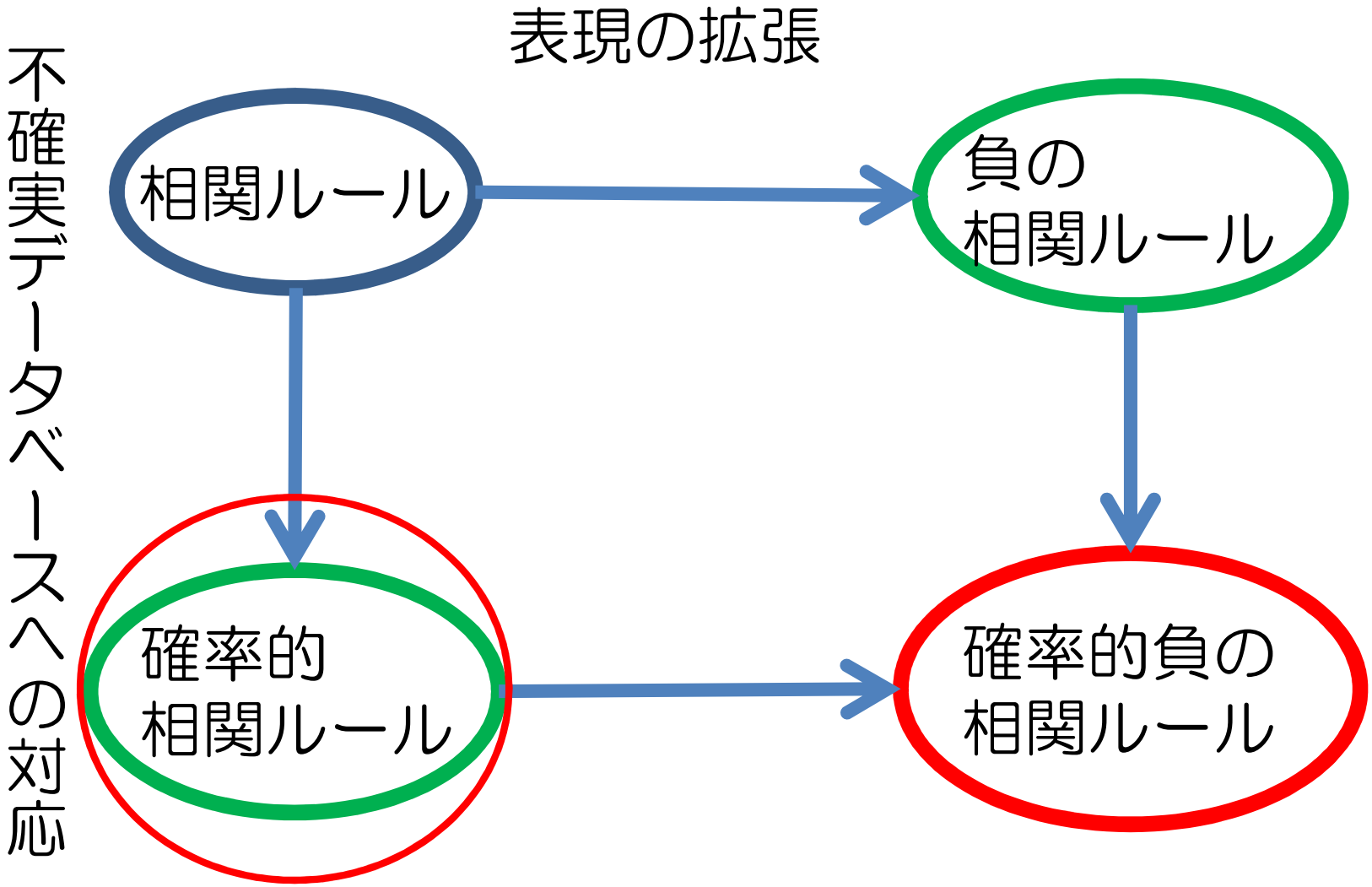
$$conf(X \Rightarrow \neg Y) \geq \mathtt{mc}$$

$$sup(X\,\neg Y) \geq \mathtt{ms}$$

$$sup(X) \geq \mathtt{ms} \wedge sup(Y) \geq \mathtt{ms}$$

$$sup(XY) < \mathtt{ms}$$

mc(最小確信度),ms(最少支持度)は
ユーザーが任意に定める閾値


表現の拡張
不確実データベースへの対応
相関ルール
負の
相関ルール
確率的
相関ルール
確率的負の
相関ルール

確率的相関ルール

確率的相関ルールとは不確実データベースを対象とした相関ルール

What is 不確実データベース？

データが存在するか否かの確率(存在確率)を持つデータベース

不確実データベースの種類

アイテムに基づく不確実データベース

ID	アイテムセット
T_1	{イチゴ:24%,トマト:33%}
T_2	{イチゴ:11%,バナナ:27%,みかん:76%}

トランザクションに基づく不確実データベース

ID	アイテムセット	%
T_1	{イチゴ,トマト,バナナ,ブドウ,リンゴ}	24
T_2	{イチゴ,バナナ,みかん,リンゴ}	47
T_3	{トマト,バナナ,みかん}	29

可能世界意味論

可能世界とは？

- 可能世界意味論(possible world semantics)

- 存在確率に従い複数の世界を考える
  - 各データが存在する世界と存在しない世界に分ける
  - 各世界が確率的に存在する

可能世界による場合分け

アイ

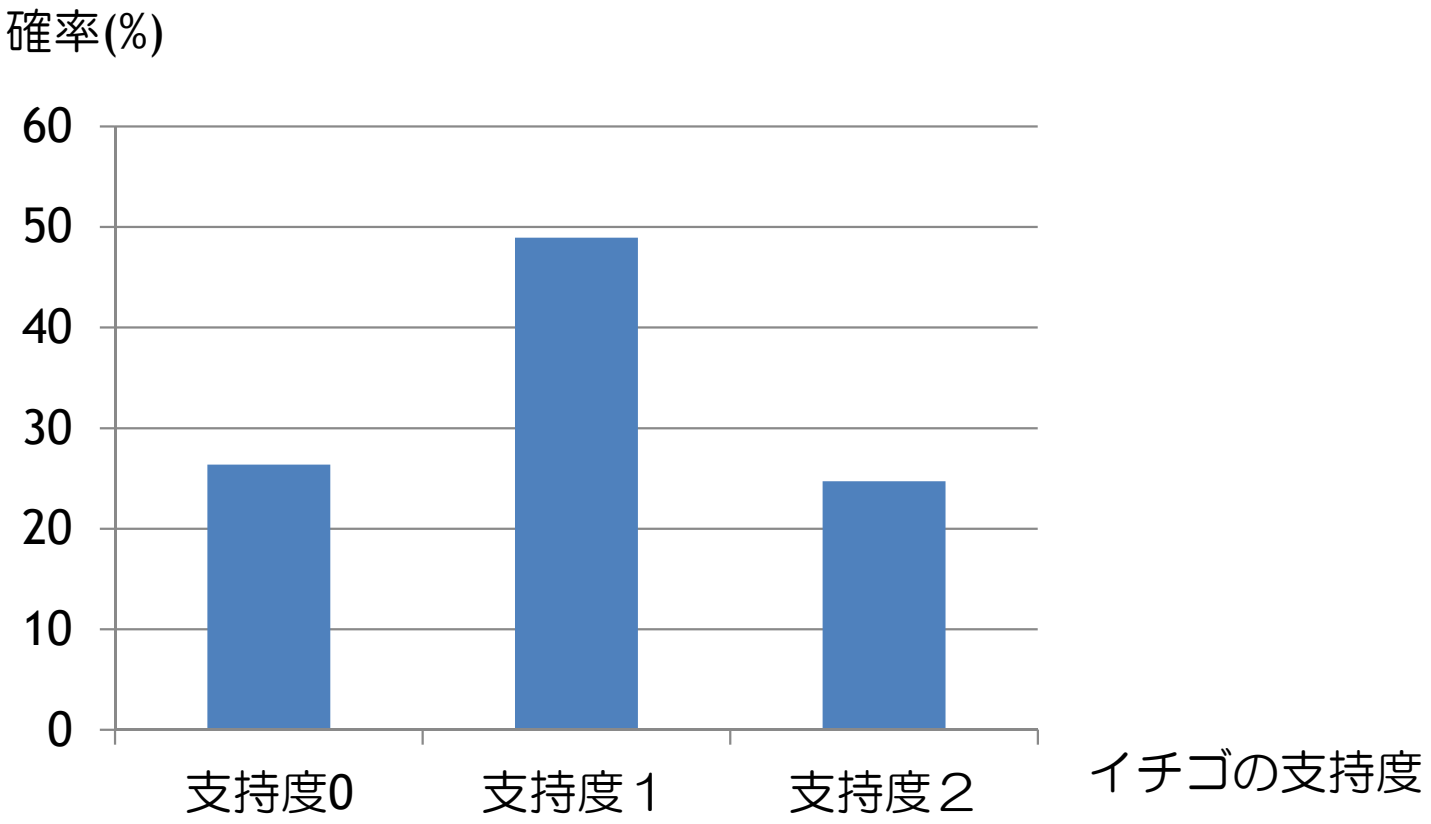
W

ID	T_1	T_2	確率 (%)
w_1	{ イチゴ, トマト }	{ イチゴ, みかん }	6.67
w_2	{ イチゴ, トマト }	{ イチゴ }	2.59
w_3	{ イチゴ, トマト }	{ みかん }	7.83
w_4	{ イチゴ, トマト }	{ }	7.83
w_5	{ イチゴ }	{ イチゴ, みかん }	14.18
w_6	{ イチゴ }	{ イチゴ }	5.51
w_7	{ イチゴ }	{ みかん }	12.76
w_8	{ イチゴ }	{ }	4.07
w_9	{ トマト }	{ イチゴ, みかん }	3.92
w_10	{ トマト }	{ イチゴ }	1.52
w_11	{ トマト }	{ みかん }	4.60
w_12	{ トマト }	{ }	1.79
w_13	{ }	{ イチゴ, みかん }	8.33
w_14	{ }	{ イチゴ }	2.66
w_15	{ }	{ みかん }	9.78
w_16	{ }	{ }	3.24

イチゴの支持度は？

ID	T_1	T_2	確率 (%)	
w_1	{ イチゴ, トマト }	{ イチゴ, みかん }	6.67	sup(イチゴ)=2
w_2	{ イチゴ, トマト }	{ イチゴ }	2.59	
w_3	{ イチゴ, トマト }	{ みかん }	7.83	sup(イチゴ)=1
w_4	{ イチゴ, トマト }	{ }	7.83	
w_5	{ イチゴ }	{ イチゴ, みかん }	14.18	
w_6	{ イチゴ }	{ イチゴ }	5.51	
w_7	{ イチゴ }	{ みかん }	12.76	
w_8	{ イチゴ }	{ }	4.07	
w_9	{ トマト }	{ イチゴ, みかん }	3.92	
w_10	{ トマト }	{ イチゴ }	1.52	
w_11	{ トマト }	{ みかん }	4.60	sup(イチゴ)=0
w_12	{ トマト }	{ }	1.79	
w_13	{ }	{ イチゴ, みかん }	8.33	
w_14	{ }	{ イチゴ }	2.66	
w_15	{ }	{ みかん }	9.78	
w_16	{ }	{ }	3.24	

評価尺度



支持度は確率変数となる

評価尺度

パターンXが支持度i以上である確率

$$P_{\geq i}(X) = \sum_{w_j \in W, (S(X, w_j) \geq i)} P(w_j)$$

S(X,wj)は世界wjにおけるパターンXの支持度

評価尺度

$$P_{\geq \text{ms}}(X) = \sum_{w_j \in W, (S(X, w_j) \geq \text{ms})} P(w_j) \quad \geq \text{mp}$$

ユーザーの定めた閾値ms,mpを満たすものを
確率的頻出パターンと言う

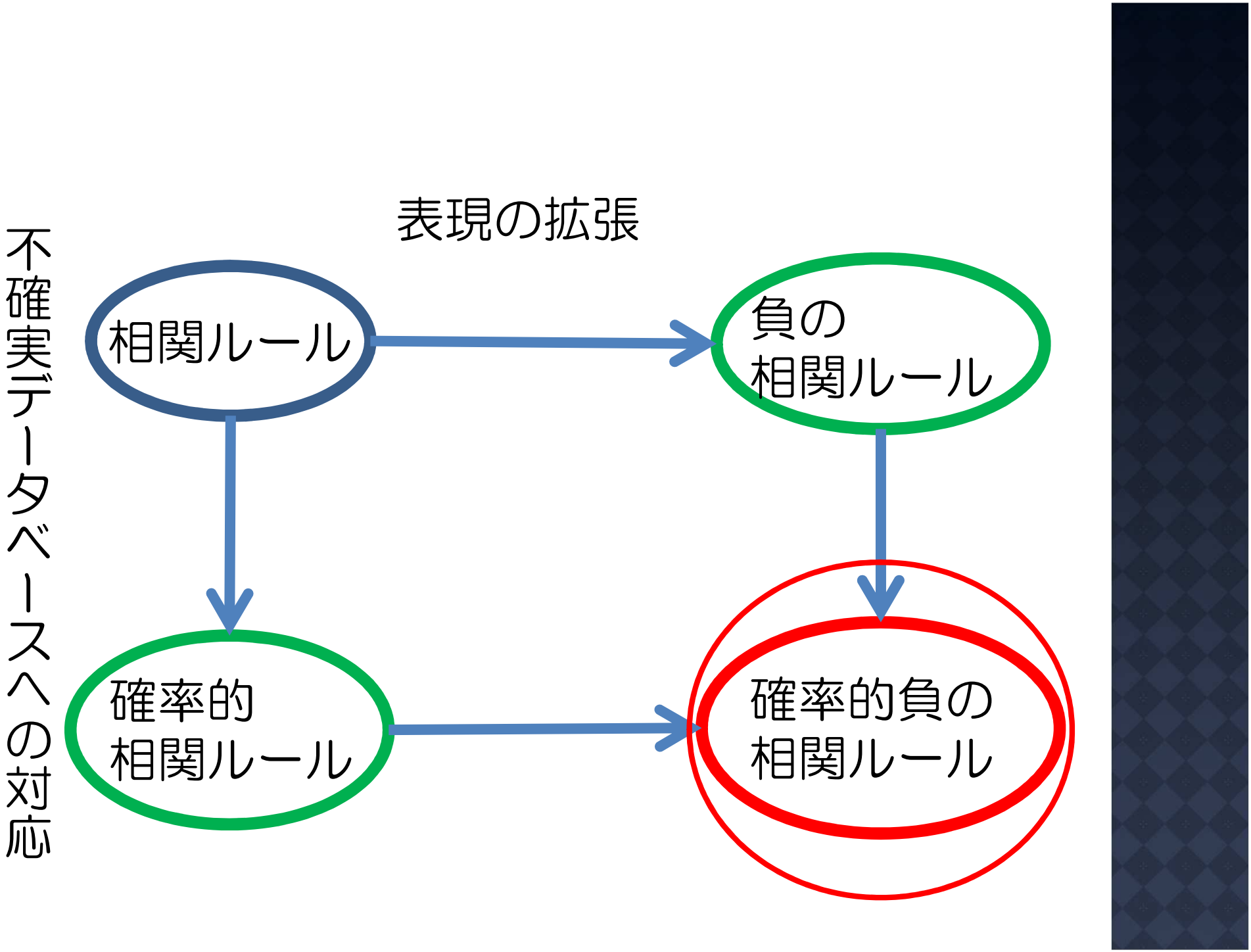
確率的相関ルール

普通の相関ルールは...

支持度同様に確率を考慮すると...

$$P\left[\begin{array}{l} conf(XY) \geq \mathrm{mc} \\ sup(XY) \geq \mathrm{ms} \wedge \end{array}\right] = \sum_{w_j \text{は条件を満たす世界}} P(w_j)$$

可能世界に展開することで計算可能


表現の拡張
不確実データベースへの対応
相関ルール
負の
相関ルール
確率的
相関ルール
確率的負の
相関ルール

確率的負の相関ルール

同様に負の相関ルールは...

$$P\left[\begin{array}{l} sup(X \Rightarrow \neg Y) \geq \mathrm{ms} \wedge \\ conf(X \Rightarrow \neg Y) \geq \mathrm{mc} \wedge \\ sup(X) \geq \mathrm{ms} \wedge \\ sup(Y) \geq \mathrm{ms} \wedge \\ sup(XY) < \mathrm{ms} \end{array}\right] = \sum_{w_j \text{は条件を満たす世界}} P(w_j)$$

原理的には可能世界に展開することで計算可能

確率的負の相関ルール

この定義を満たすもの計算するために
式変形を行う

$$P(X \Rightarrow \neg Y) = \sum_{i=\mathrm{ms}}^{N} P_i(X\neg Y) \sum_{j=0}^{\min\left(\mathrm{ms}-1, \frac{(1-\mathrm{mc})\cdot i}{\mathrm{mc}}\right)} P_j(XY) \sum_{k=\mathrm{ms}-j}^{N} P_k(\neg XY)$$

既存研究を用いて可能世界に展開することなく
計算可能！

探索の方針

$$P\left[\begin{array}{l} sup(X \Rightarrow \neg Y) \geq \mathrm{ms} \wedge \\ conf(X \Rightarrow \neg Y) \geq \mathrm{mc} \wedge \\ sup(X) \geq \mathrm{ms} \wedge \\ sup(Y) \geq \mathrm{ms} \wedge \\ sup(XY) < \mathrm{ms} \end{array}\right]$$

定義の一部に着目し前件,後件の条件に

$$P_{\geq \mathrm{ms}}(X) \geq \mathrm{mp}$$

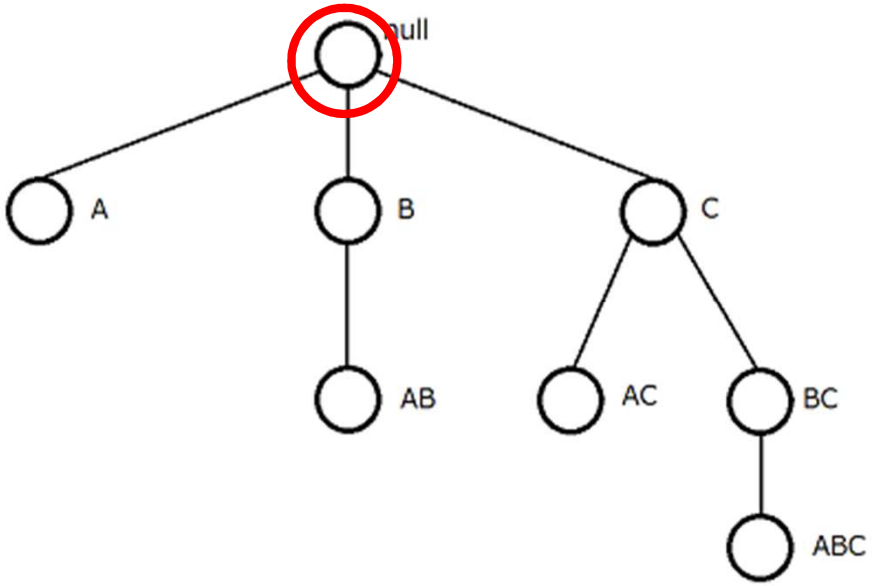
$$P_{\geq \mathrm{ms}}(Y) \geq \mathrm{mp}$$

前件,後件共に確率的頻出なパターンのみを対象にする.

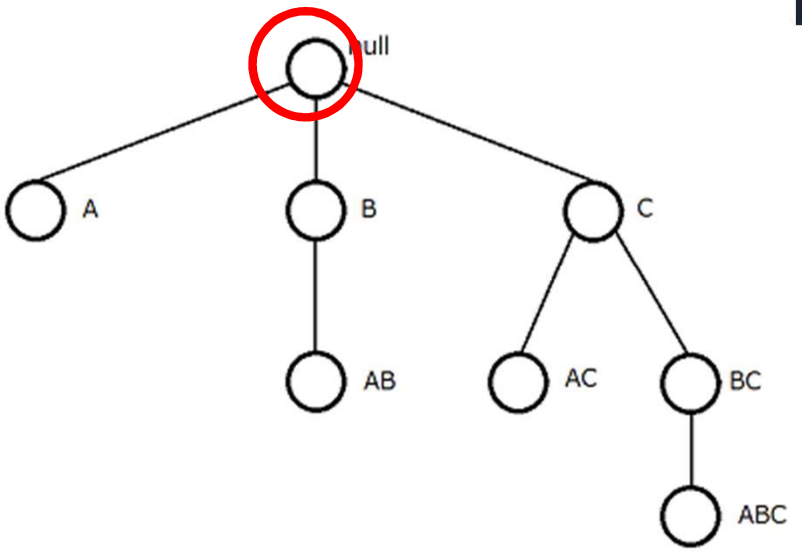
木構造の探索

- 既存手法により確率的頻出パターンを抽出
- 抽出されたパターンを木に配置

前件



後件



実験結果

- Java言語を用いて実装
- 利用したデータ
  - Frequent Itemset Mining Dataset Repository から入手したretail データ

    1000件(ms:10,mp:0.2,mc:0.2),227個,1997秒
  - twitterより入手したツイートのデータ

    1000件(ms:10,mp:0.2,mc:0.2),607個,7987秒
  - TaFengDataset より入手した買い物データ

    1000件(ms:5,mp:0.2,mc:0.2),8632個,25382秒

実験結果

$$X \Rightarrow \neg Y$$

前件		後件
フォロー	➡	ます,おはよう
楽天	➡	裏ワザ
楽天,本	➡	私
相互,初心者	➡	♪
月,年	➡	おはよう,ます

まとめ

- 不確実データベースから確率的負の相関ルールの抽出方法を提案した
- 提案に沿って実装をし,実験を行い結果を得た

今後の課題

もう１つの形式の不確実データベースに対して
同様の計算及び実装や,実装面では並列計算による
高速化が可能であるかを検証が必要.