Ann. Data. Sci. (2014) 1(1):25–39
DOI 10.1007/s40745-014-0011-y

Proposal of New Objective Measures for Mining Association Rules: Cannibalization and Unexpectedness

Hidenobu Hashikami · Masato Koda

Received: 1 September 2012 / Revised: 31 December 2012 / Accepted: 20 April 2013 /
Published online: 27 May 2014

Abstract In view of the problems that a few of the existing measures for association rules does not directly meet user's requirements, and association mining algorithms produce huge number of trivial rules, this paper proposes two new objective measures for mining association rules to solve the problems. The first measure is the degree of cannibalization between itemsets, which is bounded up with marketing strategy, and the second is the objective measure that intends to discover unexpected rules in the database. Experimental studies with application to public dataset and comparison of running time using synthetic datasets demonstrate the validity and effectiveness of the proposed measures.

1 Introduction

Association rules, i.e., a class of important regularities discovered in databases, have proven very useful in practical applications [1]. However, mining techniques of association rules still need to solve three important problems, (1) development of faster algorithms; (2) generation of appropriate measures that better fit user's specific requirements [2]; and (3) development of efficient algorithms to reduce huge number of redundant rules, most of which are of no practical interest [3]. First problem has been solved by the development of advanced algorithms, e.g., Apriori, AprioriTid [4,5],

H. Hashikami (✉)
Ricoh Company, Ltd., Ebina, Japan
e-mail: hidenobu.hh.hashikami@nts.ricoh.co.jp

M. Koda
University of Tsukuba, Tsukuba Science City, Japan
e-mail: koda@sk.tsukuba.ac.jp

and Eclat [6]. Hence this study focuses on the second and third problems because they heavily influence the quality of outputted association rules.

Since their introduction in 1993 by Argawal et al. [1], numerous measures and algorithms have been developed to deal with the frequent itemsets to solve mining problems related with association rules. Popular objective measures are *support*, *confidence*, *actionability* [7], *all-confidence* [8], and *interest* (often called *lift*) [9]. Other objective measures are proposed and analyzed in [2,10], where algorithms to select the right measure from existing measures are presented to solve the second problem. Most of the existing measures, however, do not 'directly' lead to user's goal objectives such as marketing strategy, promotional pricing, or product placements: Users have to determine the most appropriate measures that best fit their aims and objectives expected from association analysis.

In this paper, as a measure to be used for creating marketing strategy, we investigate and propose a new objective measure appropriate for mining cannibalization between itemsets in a database of sales transactions. Cannibalization in marketing strategy is defined as the negative impact of a company's new product on the sales performance of its existing related products [11]. For example, when Ricoh puts out a new printer in the market, it is a major concern that older printers will suffer from possible erosion of sales or market share; that erosion is referred to as cannibalization. There has been considerable research on cannibalization in marketing arena which can lead to a meaningful reduction of the actual and potential value of investments. Some recent studies reported the cannibalization and synergistic effects found in online-offline multichannel systems [12], and a comparison was made in the innovation similarities and differences of new product and service [12]. Cannibalization is a very real threat for the vast majority of new product launches, and becomes important topics of research. The proposed measure in this paper is the degree of cannibalization between itemsets in database, and is defined by applying dissimilarity index. As an example, 2.5 quintillion bytes of data are created in every day (see http://www-01.ibm.com/software/data/bigdata/ for detail), and data analysts are expected to play an important role in preparing an investment strategy. The appropriate measure which enables to discover cannibalizing items with evaluated value in the database, such as point-of-sale data and online shopping data, data analysts can make a precise move by identifying cannibalizing items in the database.

In order to solve the third problem that association mining algorithms tend to produce huge number of redundant rules, most of which are of no interest, we propose a new objective measure to discover unexpected rules in the database. Ref.[3] proposes somewhat subjective interestingness that leverages the user's existing domain knowledge to solve the third problem. In this paper, however, we propose an objective measure defined through designing a three dimensional in-out model, where inputs are conditional probability of one set given another set and similarity index, and output is the new measure proposed. The use of the two independent measures as inputs can enhance the capability of discovering interesting patterns hidden in huge number of rules. The effectiveness of the proposed measures is verified through the experiments applied to the public dataset and the experiment on running time using synthetic datasets.

This paper is organized as follows. In Sect. 2, we review the Apriori algorithms. In Sect. 3, we propose new objective measures, and association-rule-mining algorithm is presented in Sect. 4. Section 5 describes verification of database applications and comparison of computational cost, and Sect. 6 presents the conclusion and remarks on future studies.

2 Apriori Algorithm

Apriori is a standard algorithm to discover association rules between items in a large database of sales transactions. Formal statement of the problem is as follows [1,4,5]:.

Let $\mathcal{I}$ be a set of literals, called items. Let $\mathcal{T}$ be database of transactions, where each transaction $\mathfrak{t} \in \mathcal{T}$ is a set of items such that $\mathfrak{t} \subseteq \mathcal{I}$. In other words, $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$ be a set of items over the binary domain $\{0, 1\}$. A tuple $\mathfrak{t}$ of the database $\mathcal{T}$ is represented by identifying the items with value 1. Associated with each transaction is a unique identifier, called its *TID*. An itemset $X \subset \mathcal{I}$ is called a *k-itemset* if it contains k items. We say that a transaction $\mathfrak{t}$ contains an itemset X, if $X \subseteq \mathfrak{t}$. For X, a maps $'$ is defined to obtain a set of transactions common to all the items in X as follows:

$$X' := \{\mathfrak{t} \in \mathcal{T} \mid X \subseteq \mathfrak{t}\}, \tag{1}$$

The maps $' : X \mapsto X'$ is traditionally called *polars* [14].

An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. X is referred to as the antecedent of the rule, and Y as the consequent. The rule $X \Rightarrow Y$ has *support s* in the database of transactions $\mathcal{T}$ if $s\%$ of transactions in $\mathcal{T}$ satisfy $X \cup Y$. The support is defined as:

$$support(X \Rightarrow Y) = \frac{|(X \cup Y)'|}{|\mathcal{T}|} \in [0, 1], \tag{2}$$

where symbol $|\cdot|$ represents the number of elements, thus, $|(X \cup Y)'|$ is the number of transactions that contain common items in X and Y, and $|\mathcal{T}|$ is the total number of transactions. The support is defined to be the fraction of transactions in $\mathcal{T}$ that satisfy the union of items in X and Y of the rule.

The rule $X \Rightarrow Y$ holds in $\mathcal{T}$ with *confidence c* if $c\%$ of transactions in $\mathcal{T}$ that satisfy X also satisfy Y. The confidence is defined as:

$$confidence(X \Rightarrow Y) = \frac{|(X \cup Y)'|}{|X'|} \in [0, 1], \tag{3}$$

where $|X'|$ is the number of transactions that contain all items in X. The $confidence(X \Rightarrow Y)$ is denoted as $conf$ in this paper. The confidence for a rule is the conditional probability of having Y contained in a transaction, given that X is contained in that transaction. Here, a *lift* is used as one measure of the predictive strength for a rule, however, we leave out the explanation of the lift because it is unrelated to this study.

Because the databases involved in the applications are very large, Ref. [4] presented fast algorithms for mining association rules, called *Apriori*, etc. Given the database, the apriori algorithm finds association rules that have support and confidence greater than the user-specified minimum values, respectively.

3 Proposal of New Measures

In this section, new objective measures for mining association rules, called *cannibalization* and *unexpectedness*, are presented. The *cannibalization* measures the degree of cannibalization between itemsets, and is defined by applying the Bray-Curtis dissimilarity which is a statistic to compare the dissimilarity of sets used in ecology and biology. The *unexpectedness* is intended to discover unexpected association rules from the database, and is defined through the conditional probability and the Sørensen-Dice coefficient which is one of similarity indices. As basic strategy, a three dimensional in-out model is designed, where the inputs are confidence and similarity index and the output is the new measure proposed. In the following, we introduce the coefficient used in this study, and describe the proposed measures.

3.1 Coefficient of Similarity and Dissimilarity

In the proposed measures, we use the Bray-Curtis dissimilarity which is a statistic used to quantify the compositional dissimilarity of dichotomous variables (two and only two categories (e.g., Yes or No, True or False)) introduced by Bray and Curt in [15].

Let Z be a 2×2 contingency table, see Table 1. Z is composed of two dichotomous variables taken from X and Y, each with a binary attribute $\{0, 1\}$ representing purchase information. Each attribute of X and Y can either be 0 or 1. If $X \subseteq \mathrm{t}$ for $\mathrm{t} \in \mathcal{T}$ then attribute of X is 1 and if $X \nsubseteq \mathrm{t}$ then 0, thus, 1 means all item of X are bought and 0 means some items of X are not. Let n_{11}, n_{10}, n_{01}, and n_{00} denote:

$$\begin{aligned} n_{11} &= \mid \{\mathrm{t} \in \mathcal{T} \mid x \in \mathrm{t} \wedge y \in \mathrm{t} : \forall x \in X, \forall y \in Y\} \mid \\ &= \mid \{\mathrm{t} \in \mathcal{T} \mid (X \cup Y) \subseteq \mathrm{t}\} \mid, \\ n_{10} &= \mid \{\mathrm{t} \in \mathcal{T} \mid x \in \mathrm{t} \wedge y \notin \mathrm{t} : \forall x \in X, \forall y \in Y\} \mid, \\ n_{01} &= \mid \{\mathrm{t} \in \mathcal{T} \mid x \notin \mathrm{t} \wedge y \in \mathrm{t} : \forall x \in X, \forall y \in Y\} \mid, \\ n_{00} &= \mid \{\mathrm{t} \in \mathcal{T} \mid x \notin \mathrm{t} \wedge y \notin \mathrm{t} : \forall x \in X, \forall y \in Y\} \mid, \end{aligned}$$

Table 1 Elements of the 2 × 2 contingency table

X	Y		
	1	0	
1	n_{11}	n_{10}	n_{1Y}
0	n_{01}	n_{00}	n_{0Y}
	n_{X1}	n_{X0}	$\mid \mathcal{T} \mid$

respectively. In Table 1, n_{11}, n_{10}, n_{01}, and n_{00} are the number of transactions satisfied with above condition, for example, n_{10} represents the number of transactions that all items of X are bought but any items of Y are not. Let n_{X1} be the sum of the first row and $\cdot$ equals $n_{11} + n_{01}$, similarly, $n_{X0} = n_{10} + n_{00}$, $n_{1Y} = n_{11} + n_{10}$, and $n_{0Y} = n_{01} + n_{00}$. The formula for computing the Bray-Curtis dissimilarity $BC(X, Y)$ of X and Y is given as:

$$BC(X, Y) = \frac{n_{10} + n_{01}}{n_{1Y} + n_{X1}} = \frac{n_{10} + n_{01}}{2n_{11} + n_{10} + n_{01}} \in [0, 1]. \tag{4}$$

$BC(X, Y)$ ranges from 0 to 1, where 0 means the two itemsets have the same composition (that is they share all items), and 1 means the two itemsets do not share any items. The Bray-Curtis dissimilarity directly related to the Sørensen-Dice coefficient $SD(X, Y)$ between the same itemsets: $SD(X, Y) = 1 - BC(X, Y) \in [0, 1]$ [16,17]. It is a statistic for comparing the similarity of two itemsets.

3.2 The *Cannibalization*

A challenge in association analysis is to provide the appropriate measure that best fit user's requirements. For our purpose of bounding up with marketing strategy directly, the *cannibalization* is proposed as a novel quality measure for mining cannibalization between itemsets in the database. The *cannibalization* is based on the Bray-Curtis dissimilarity. We first explain a phenomenon of cannibalization commonly occurs in the database.

Cannibalization is a real threat for the vast majority of new product launches, which is defined as the negative impact of a company's new product on the sales performance of its existing related products [11]. In the database of transactions $\mathcal{T}$, when an item of X (new product) cannibalizes Y (existing products), it is rare that items of X and Y is bought simultaneously. Thus, cannibalization between X and Y in $\mathcal{T}$ represents a frequent occurrence of $x \in \mathrm{t} \wedge y \notin \mathrm{t}$ and $x \notin \mathrm{t} \wedge y \in \mathrm{t}$, corresponding to events of n_{10} and n_{01} (the readers are referred to Table 1). If these are extremely occurred in $\mathcal{T}$, the Bray-Curtis dissimilarity of X and Y approaches the maximum value of 1. The reasons for use of the Bray-Curtis dissimilarity are because purchase information in which X and Y are not bought simultaneously (n_{00}) need not be considered in Bray-Curtis, and the denominator of Bray-Curtis represents the number of gross sales for items of X and Y ($n_{1Y} + n_{X1} = 2n_{11} + n_{10} + n_{01}$). There seems to be no alternative similarity and dissimilarity index.

Bray-Curtis dissimilarity essentially denotes that the ratio of the number of cannibalizing items to the number of gross sales for itemsets. Hence, for an association rule $X \Rightarrow Y$, $cannibalization(X \Rightarrow Y)$ is defined as:

$$\begin{aligned} &cannibalization(X \Rightarrow Y) \\ &\quad = BC(X, Y) = \frac{n_{10} + n_{01}}{2n_{11} + n_{10} + n_{01}} \in [0, 1]. \end{aligned} \tag{5}$$

It is a symmetrical definition and the value ranges from 0 to 1, where 0 means the two itemsets share the same items (that is they do not cannibalize), and 1 means that the two itemsets cannibalize. A degree of cannibalizing item X also writes $\frac{n_{10}}{2n_{11}+n_{10}+n_{01}} \in [0, 1]$, and a degree of cannibalized item Y writes $\frac{n_{01}}{2n_{11}+n_{10}+n_{01}} \in [0, 1]$. Many association rules may be biased toward lower *cannibalization*, and hence a crawling algorithm may find association rules that satisfy $support(X)$ and $support(Y)$ greater than the user-specified minimum support. Moreover, since the *cannibalization* only deal with the products belonging to the same category, cannibalizing products can be included in a same market.

3.3 The *Unexpectedness*

The number of derived association rules is represented by $\sum_{k=2}^{m} {}_mC_k(2^k - 2)$ provided m-itemset, for example, 57002 rules are generated if $m = 10$. The number of association rules grow exponentially huge as the number of the itemset becomes larger. The support, confidence, and lift used in Apriori algorithm can find huge number of association rules, however, they can only give trivial rules taken from the top-ranked association rules. To resolve this problem, we propose the *unexpectedness* as a novel quality measure to discover unexpected association rules in the database. It is abstractly defined as 'the items purchased simultaneously although they are not directly correlated or unrelated.' For example, at a first glance it seems like there is no relation between beer and diapers, but there are numerous cases where the two items are bought simultaneously in high frequency. To discover seemingly unrelated but simultaneous purchased items, we represent the relation between items using the Sørensen-Dice coefficient and define confidence by simultaneous purchase frequency.

To capture the *unexpectedness* for an association rule, a three dimensional in-out model is developed, where inputs are the Sørensen-Dice coefficient and the confidence, and output is *unexpectedness*. Fig.1 (left) depicts the model. The model is defined by the condition of maximum output value 1 when $SD(X, Y) = 1$ and $conf = 1$ (hereafter referred to as the maximum end point), while the other end points are 0. When the function met the described condition, we use logical product of t-norm

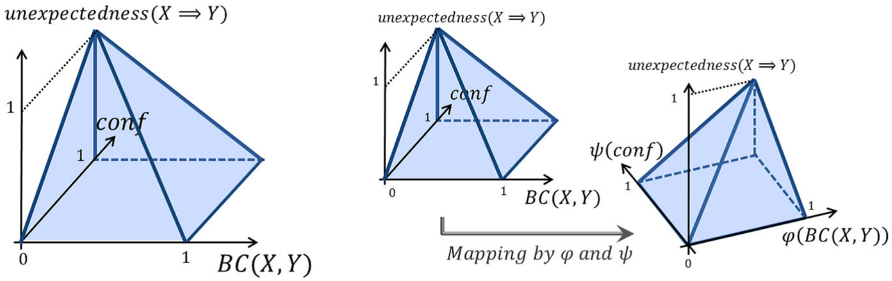
$unexpectedness(X \Rightarrow Y)$
1
$conf$
1
0
1
$BC(X,Y)$
$unexpectedness(X \Rightarrow Y)$
1
$conf$
1
0
1
$BC(X,Y)$
Mapping by φ and ψ
$unexpectedness(X \Rightarrow Y)$
1
$\psi(conf)$
1
1
$\varphi(BC(X,Y))$
0

Fig. 1 (*Left*) is In-out model of unexpectedness for an association rule (logical product of t-norm). (*Right*) is conceptual diagram of the mappings by φ and ψ

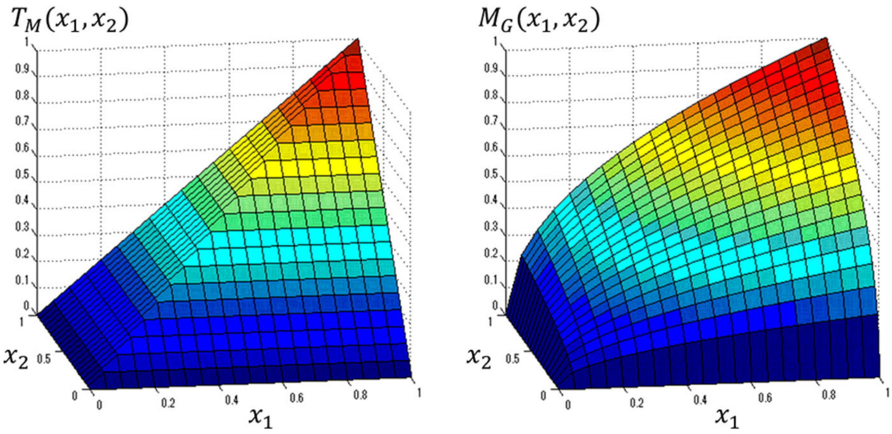

Fig. 2 Three dimensional in-out model of logical product of t-norm and geometric mean of averaging operator

(unabbreviated, triangular norm) and geometric mean of averaging operator which are binary operations used in probabilistic metric spaces or fuzzy logic [18,19]. We call them t-norm and λ-sum for short. Fig.2 shows three dimensional graphs of them. Let T_M be t-norm and M_G be λ-sum, where $T_M : [0, 1]^2 \rightarrow [0, 1]$ and $M_G : [0, 1]^2 \rightarrow [0, 1]$, such that $\forall x_1, x_2 \in [0, 1]$. T_M and M_G are given by $T_M(x_1, x_2) = \min\{x_1, x_2\}$ and $M_G(x_1, x_2) = \sqrt{x_1 x_2}$, respectively. The $unexpectedness(X \Rightarrow Y)$ by T_M and M_G are defined as:

$$\begin{aligned} & unexpectedness(X \Rightarrow Y) \\ & \quad = \min\{\varphi(SD(X, Y)), \psi(conf)\} \quad (t - norm), && (6) \\ & \quad = \sqrt{\varphi(SD(X, Y))\psi(conf)} \qquad (\lambda - sum), && (7) \end{aligned}$$

where

$$\varphi(SD(X, Y)) = 1 - SD(X, Y) \in [0, 1], \tag{8}$$

$$\psi(conf) = conf \in [0, 1]. \tag{9}$$

The $\varphi(SD(X, Y))$ is a mapping that linearly transforms $SD(X, Y) \in [0, 1]$ with the transportation function $\min\{\varphi(0), \psi(1)\} = 1$ ($\sqrt{\varphi(0)\psi(1)} = 1$), that is, a value of the maximum end point becomes 1 in space of t-norm or λ-sum. Similarly, $\psi(conf)$ is a mapping that linearly transforms conf $\in [0, 1]$. Fig. 1 (right) shows conceptual diagram of the mappings.

3.4 Examples of New Measures

Consider the database of transactions given in Table 2 over the set of items $\mathcal{I} = \{$desktop computer, laptop, netbook, tablet, e-book reader$\}$.

Table 2 An example database of transactions

TID	Items
1	Desktop computer, laptop, tablet
2	Desktop computer, laptop, e-book reader
3	Laptop, e-book reader
4	Desktop computer, netbook, tablet
5	Desktop computer, laptop
6	Netbook, tablet
7	Desktop computer, netbook, e-book reader
8	Desktop computer, laptop
9	Desktop computer, laptop, tablet, e-book reader

Table 3 Elements of the 2 × 2 contingency table for *cannibalization* and *unexpectedness*

cannibalization		{e-book reader}		
		1	0	
	1	1	3	4
{tablet}	0	3	2	5
		4	5	9
unexpectedness		{desktop computer}		
		1	0	
	1	2	1	3
{e-book reader, laptop}	0	5	1	6
		7	2	9

The classic algorithm, for instance, derives an association rule [laptop] ⇒ [desktop computer] with relatively high support 55.6 % (5/9) and confidence 83.3 % (5/6) as the association rule with relatively high support and confidence. In this subsection, we give examples of association rules for proposed measures. The *cannibalization* for a rule [tablet] ⇒ [e-book reader] is derived. The rule has support 11.1 % (1/9) and confidence 25.0 % (1/4). Table 3 (top) shows a fourfold table for the rule, and then the number of gross sales for tablet and e-book reader is 8. The *cannibalization* is $6/8 = 0.75$ (75.0 %). As shown in Table 2, both of tablet and e-book reader are not bought simultaneously in most transactions. The *unexpectedness* for a rule [e-book reader, laptop] ⇒ [desktop computer] is derived. The rule has support 22.2 % (2/9) and confidence 66.7 % (2/3). Table 3 (bottom) shows a fourfold table for the rule, and then the value of the Sørensen-Dice coefficient is $4/10 = 0.4$. The *unexpectedness* by t-norm is $\min\{1 - 0.4, 0.667\} = 0.6$ (60.0 %). This is most unpredictable rule in the database.

4 Rules Generation Algorithm

This section exolores an algorithm to generate association rules for new indicators. Basic framework of the algorithm is based on the *Apriori* algorithm to discover all association rules in large itemsets [4]. The algorithm to solve the association mining problem is divided into two phases. In the first phase, all frequent itemsets are generated. The second phase consists of the generation of all support and confident association rules.

Algorithm 1 `Apriori`($\mathcal{T}$,t_{sup}).

```
$L_1 \leftarrow \{i \in \mathcal{I} \mid support(i) \geq t_{sup}, \forall i \in \mathcal{I}\}$;
$k \leftarrow 2$;
    while $L_{k-1} \neq \emptyset$ do
        $C_k \leftarrow$ apriori − gen $(L_{k-1})$;
        foreach t in $\mathcal{T}$ do
            foreach $c \subseteq$ t in $C_t$ do
                $c.count \leftarrow c.count + 1$;
        if $name\_ofmeasure =$ "$U$" then
            $L_k \leftarrow \{c \in C_k \mid support(c) \geq t_{sup}, \forall c \in C_k\}$;
        if $name\_of\_measure =$ "$C$" then
            $L_k \leftarrow \{c \in C_k \mid support(c) < t_{sup}, \forall c \in C_k\}$;
        $k \leftarrow k + 1$;
$L \leftarrow \bigcup_k L_k$ ;

Subroutine apriori − gen $(L_{k-1})$
$C_k \leftarrow \emptyset$;
foreach $c_p, c_q \in L_{k-1} \mid c_p[i] = c_q[i], \forall i \in \{1, \ldots, k-2\}$ do
    $c \leftarrow c_p \cup c_q = \{c_p[1] \ldots, c_p[k-2], c_p[k-1], c_q[k-1]\}$;
    if $l_{k-1} \subset c \mid \forall l_{k-1} \in L_{k-1}$ then
        $C_k \leftarrow C_k \cup c$;
return $C_k$;
```

Assume for simplicity that items in transactions and itemsets are kept sorted in their lexicographic order. Associated with each itemset is a count field to store the support for this itemset. We say that L_k is a set of k-itemsets l_k with items and support count, and C_k is a set of candidate k-itemsets. Instead of flags of *cannibalization* and *unexpectedness*, we use "C" and "U" in the pseudocode of algorithm for short.

In the algorithm, there are two procedures, `Apriori`($\mathcal{T}$,t_{sup}) and `Ap-genrules` ($\mathcal{T}$,L,t_{sup},t_{conf},t_{bc},t_{sd}), where t_{sup}, t_{conf}, and t_{bc} are threshold values of support, confidence, Bray-Curtis dissimilarity, and Sørensen-Dice coefficient defined by user, respectively. `Apriori` is an algorithm for mining itemset. `Ap-genrules` is an algorithm for the generation the rules and runs after `Apriori`.

The pseudocode is presented below, with line-by-line explanation of each procedure.

Algorithm 2 `Ap-genrules`$(\mathcal{T},L,t_{sup},t_{conf},t_{bc},t_{sd})$.

```
foreach $l_k$ in $L_k, k \geq 2$ do
    $\_s_k \leftarrow$ subsets $(l_k)$;
    foreach $b_k$ in $\_s_k$ do
        $h_k \leftarrow l_k \setminus b_k$;   # $h_k \neq \emptyset$
        if $support(b_k) \geq t_{sup}$ and $support(h_k) \geq t_{sup}$ then
            $confidence \leftarrow support(l_k)/support(b_k)$;
            $indicator \leftarrow CompMeasure\ (\mathcal{T}, confidence, b_k, k_k)$;
            if not $meaure = None$ then
                output $b_k \Rightarrow h_k$
                        $with\ measure, confidence, and\ support(l_k)$

Subroutine $CompMeasure(\mathcal{T}, confidence, b_k, k_k)$
if $name\_of\_measure =$ "$C$" then
        $bc_k \leftarrow$ CompBrayCurtis $(\mathcal{T}, b_k, h_k)$;
        if $bc_k \geq t_{bc}$ then  return $bc_k$;
if $name\_of\_measure =$ "$U$" then
        $sd_k \leftarrow$ CompSorensenDice $(\mathcal{T}, b_k, h_k)$;
        if $confidence \geq t_{conf}$ and $sd_k < t_{sd}$ then
            return $\min\{1 - sd_k, confidence\}$;
return $None$;
```

4.1 Algorithm Apriori

The `Apriori` algorithm discovers itemset satisfying t_{sup}.

Line 1 - The first pass of the algorithm simply counts the number of occurrences of each item to determine the large 1-itemsets L_1.

Lines 2 and 3 - A subsequent pass, named pass k, consists of two phases.

Line 4 - First, the large itemsets L_{k-1} found in the $k-1$th pass are used to generate the candidate itemsets C_k, using the apriori-gen function.

Lines 5 - 7 - Next, the database of transactions $\mathcal{T}$ is scanned and the support of candidates c in C_k is counted. The c in C_k that are contained in a given transaction $\mathrm{t} \in \mathcal{T}$ is used for efficiency.

Lines 8 - 12 - If type of measure is "U", candidates c satisfying $support(\mathrm{c}) \geq t_{sup}$ are determined as elements of large k-itemsets L_k. If "C", then $support(\mathrm{c}) < t_{sup}$.

Line 13 - Return set of large itemsets L having $L_1, L_2, \ldots, L_K$.

The apriori-gen function takes as an argument L_{k-1}, the set of all $(k-1)$-itemsets. It returns the superset of the set of all k-itemsets. First, in the join step, L_{k-1} with L_{k-1} joined to obtain the superset of the final set of C_k. The union $c_p \cup c_q$ of itemsets p, $q \in L_{k-1}$ is inserted in C_k if they share their $k-2$ first items $(c_p[0] = \emptyset)$. Next, in the prune step, all itemsets $c \in C_k$ are deleted such that some $(k-1)$-subset of c is not in L_{k-1}.

4.2 Algorithm Ap-Genrules

The `Ap-genrules` algorithm generates the association rules in that a consequent h_k is allowed to have more than one item [5]. For every k-itemset l_k ($k \geq 2$), all rules $b_k \Rightarrow h_k$ ($l_k \backslash b_k$) are outputted, where the antecedent b_k is a subset of l_k, such that the confidence $support(l_k)/support(b_k)$ and coefficient of similarity and dissimilarity satisfy specified threshold values t_{conf}, t_{bc} and t_{sd}, respectively.

Line 1 - Iterate over the element l_k of the large k-itemsets L_k ($k \geq 2$) given by `Apriori`.

Line 2 - To generate all possible consequent with more than one item, compute proper subsets $\_s_k$ of l_k, where $\emptyset \notin \_s_k$. As an example, let l_3 be $\{1, 2, 3\}$. The subsets function outputs $\{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$.

Line 3 - Iterate over the antecedent $b_k \neq \emptyset$.

Line 4 - The consequent h_k is computed by a complement of b_k in l_k.

Lines 5 and 6 - Compute the confidence if supports of both b_k and h_k are at least more than t_{sup}.

Line 7 - CompMeasure function computes the proposed measures for a association rule. This pseudocode presents the measure using t-norm. The measure satisfying specified threshold values is computed.

Lines 8 and 9 - If a return value from CompMeasure function is not $None$, output a rule $b_k \Rightarrow h_k$ with values of measure, confidence, and support.

5 Experimental Evaluation

We studied the validity of the proposed measures for mining association rules applied to public dataset, and also experimented on the performance of the algorithm measured by running time using synthetic datasets.

5.1 Public Dataset Experiment

The dataset used in this experiment is *Online directory of certified businesses with a detailed profile* from the small business services in the NYC Open Data which is public data generated by New York City agencies and other City organizations (see nycopendata.socrata.com for detail). It contains a list of certified minority and woman-owned business enterprises (M/WBEs), emerging business enterprises (EBEs), and locally-based enterprises (LBEs) throughout the New York City tri-state area, and also provides detailed information on certified companies, including a brief description of their work and contact information. In the experiment, we used data of 1420 companies which have certification, ethnicity and city as attributes. Below is the list in the form *attribute name* (kinds of attributes): *certification* (MBE, WBE, EBE, LBE), *ethnicity* (Asian, Black, Hispanic, Non-minority), *city* (Bronx, Brooklyn, Flushing, Jamaica, Long Island City, New York, Staten Island, Yonkers).

Table 4 Association rules for public dataset. Rules are written in the form $X \Rightarrow Y$ $(proposal, confidence, support)$

Cannibalization$(X \Rightarrow Y)$
[MBE] ⇒ [Non-minority] (99.9, 0.1, 0.1)
[MBE, Hispanic] ⇒ [Brooklyn] (92.0, 7.7, 1.3)
[Asian] ⇒ [WBE, New York] (91.7, 7.7, 1.5)
[Hispanic] ⇒ [WBE, New York] (91.7, 8.6, 1.4)
[Black] ⇒ [WBE,New York] (90.8, 7.3, 1.4)
[WBE] ⇒ [Asian] (88.8, 8.0, 3.8)
[WBE] ⇒ [Hispanic] (85.3, 9.9, 4.7)
[Non-minority, WBE] ⇒ [Brooklyn] (84.4, 11.7, 3.5)
[MBE, Asian] ⇒ [Brooklyn] (84.0, 14.1, 2.8)
unexpectedness$(X \Rightarrow Y)$
[Non-minority] ⇒ [WBE] (87.8, 100.0, 30.0)
[Black] ⇒ [MBE] (78.7, 100.0, 30.0)
[WBE] ⇒ [Non-minority] (69.6, 62.8, 30.0)
[Asian] ⇒ [MBE] (67.3, 99.0, 20.0)
[Non-minority, New York] ⇒ [WBE] (63.0, 100.0, 11.8)
[Hispanic] ⇒ [MBE] (62.7, 100.0, 16.4)
[WBE, New York] ⇒ [Non-minority] (58.0, 67.5, 11.8)
[New York] ⇒ [WBE] (51.9, 59.4, 17.5)
[New York] ⇒ [MBE] (45.1, 57.8, 17.5)

Table 4 gives the results of association rules for public dataset. the *cannibalization* and *unexpectedness* exclude symmetric and inclusive association rules. As the threshold vales in the proposed algorithm, $t_{sup} = 15\,\%$ and $t_{bc} = 80\,\%$ if type of measure is *cannibalization*, and $t_{sup} = 15\,\%$, $t_{conf} = 50\,\%$, $t_{sd} = 75\,\%$ if type of measure is *unexpectedness*, respectively. Rules are written in the form $X \Rightarrow Y \mid (proposal, c, s)$, where c is the confidence and s is the support expressed as a percentage.

The *cannibalization* yielded a rule with 99.9 % when if an owner gets certified as a MBE then is not Non-minority. MBE is the certification defined as a business owned by American citizens of ethnic minority, and the owner of non-minority do not get MBE, therefore, this rule discovered in the dataset makes sense. If most owners also are Asian, Hispanic, or Black then type of certification is not WBE and office location is not in New York. If most owners is Hispanic or Asian with MBE then office location is not in Brooklyn. Negative relation between types of certification, ethnicity, and office location, were yielded by the *cannibalization*. The *unexpectedness* provides an association rule that if an owner is Non-minority then she gets certified as a WBE. If owners also are Black, Asian, or Hispanic, then they get certified as a MBE. As shown the order by *unexpectedness* and *confidence*, the *unexpectedness* can output association rules different from the result of confidence. Among them, the *unexpectedness* yielded rules that if office location is in New York then the type of certification is MBE or WBE.

5.2 Computational Cost

To assess the performance of the algorithm over a large operating region, synthetic transactions data was developed. These transactions attempt to mimic the transactions in the retailing environment. To create synthetic datasets, the following method was used. First 200 potentially large itemsets from 100 items were generated. Picking the size of a set from a Poisson distribution with mean equal to $| I | = 2$ or 4 we randomly assigned items to the set. To model the characteristics that large itemsets often have common items, some fraction of items in the subsequent itemsets were chosen from the previous itemset generated. Then $| D | = 100{,}000$ transactions were generated. The average size $| T |$ of a transaction was 5 or 10and the size was picked from a Poisson distribution. Each transaction was assigned a series of fractions of potentially large itemsets, to model that all the items in a large itemset are not always bought together. Table 5 summarizes the dataset parameter setting. A more detailed description of the synthetic data generation can be found in [4].

Table 6 shows the execution times for the three synthetic datasets given in Table 5 in the order of decreasing values of minimum support. As the minimum support decreases, the execution times of all the algorithms increase because of increases in the total number of candidate and large itemsets. Apriori outperforms the proposed algorithm for most of problem sizes, by factor ranging from 15 % for high minimum support to more than an order of magnitude for low levels of minimum support. This is because extra calculation `Ap-genrules` algorithm calculates proposed measures is needed.

The experiment was conducted using Intel® Core(TM) i5 2.67 GHz CPU with 3.0 GB memory size running Ubuntu (32bit), with the algorithm implemented in Python programming language.

Table 5 Parameters and sizes of datasets

Name	$\mid T \mid$	$\mid I \mid$	$\mid D \mid$	MB
T5.I2.D100K	5	2	100K	1.6
T5.I4.D100K	5	4	100K	1.5
T10.I4.D100K	10	4	100K	2.8

Table 6 Execution time of synthetic data for variable minimum support (%) (timing in seconds)

	Minimum support (%)				
T5.I2.D100K	15%	10 %	7.5 %	5 %	3.5 %
Proposal	2.1	2.6	4.3	13.0	17.5
Apriori	2.1	2.6	4.4	12.7	16.2
T5.I4.D100K	15 %	10 %	7.5 %	5 %	3.5 %
Proposal	2.0	2.4	4.4	13.0	18.0
Apriori	1.9	2.5	4.3	13.0	16.8
T10.I4.D100K	15 %	10 %	7.5 %	5 %	3.5 %
Proposal	6.4	17.8	31.8	77.7	166.4
Apriori	5.5	15.4	25.1	56.7	113.4

6 Conclusion

In view of the problems that most of the existing measures for association rules do not directly meet user's requirements, and association mining algorithms produce huge number of redundant and/or trivial rules, this paper proposed two new objective measures for mining association rules to solve the problems. The first measure was the degree of cannibalization between itemsets, which is bounded up with marketing strategy, and the second was the objective measure that intends to discover unexpected rules in the database. It is expected that the present study will help a firm to reduce the actual or potential value of investments in creating the business strategy to avoid cannibalization. Experimental studies applied to public dataset and comparative study on running time using synthetic datasets showed the effectiveness of the proposed measures. Important future work may be a development of methodology to express cannibalization rates for mined items using the database of sales transactions. In Ref. [20], a methodology is developed to estimate cannibalization rates for pioneering innovations. The development of a methodology to visualize cannibalization effects may be another important work.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, vol. 22, pp. 207–216.
2. Tan P-N, Kumar V, Srivastava J (2004) Selecting the right objective measure for association analysis. Inform. Syst. 29(4):293–313
3. Liu B, Hsu W, Chen S, Ma Y (2000) Analyzing the subjective interestingness of association rules. IEEE Intell Syst 15(5):47–55
4. Agrawal R, Srikant R (1994), Fast algorithms for mining association rules. In: Proceedings 20th International Conference on Very Large Data, Bases, pp. 487–499.
5. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. In: Advances in knowledge discovery and data mining, pp. 307–328.
6. Zaki MJ (2000) Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390
7. Silberschatz A, Tuzhilin A (1996) What makes patterns interesting in knowledge discovery systems. IEEE Trans Knowl Data Eng 8(6):970–974
8. Omiecinski E (2003) Alternative interest measures for mining associations. IEEE Trans Knowl Data Eng 15(1):57–69
9. Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: Proceedings ACM SIGMOD International Conference on Management of Data, pp. 255–264.
10. Lenca P, Meyer P, Vaillant B, Lallich S (2008) On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. Eur J Oper Res 184(2):610–626
11. Copulsky W (1976) Cannibalism in the marketplace. J Mark 40(4):103–105
12. Kollmann T, Kuckertz A, Kayser I (2012) Cannibalization or synergy? Consumers' channel selection in online-offline multichannel systems. J Retail Consum Serv 19(2):186–194
13. Nijssen EJ, Hillebrand B, Vermeulen PAM, Kemp RGM (2006) Exploring product and service innovation similarities and differences. Int J Res Mark 23(3):241–251
14. Carpineto C, Romano G (2004) Concept data analysis: theory and applications. John Wiley and Sons, Hoboken
15. Bray RJ, Curtis JT (1957) An ordination of upland forest communities of southern Wisconsin. Ecol Monog 27:325–349
16. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

17. Sorensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Vidensk Selsk Biol Skr 5:1–34
18. Gupta MM, Qi J (1991) Theory of T-norms and fuzzy inference methods. Fuzzy Sets Syst 40(3):431–445
19. Kenney JF, Keeping ES (1962) Mathematics of statistics, Part I, 3rd edn. Van Nostrand, Princeton
20. van Heerde HJ, Srinivasan S (2007) Dekimpe, "Estimating cannibalisation rates for pioneering innovations". Mark Sci 29(6):1024–1039

Hidenobu Hashikami received the master of engineering from University of Tsukuba, Ibaraki, Japan. He has been working at Ricoh Company, Ltd. since 2012. His research interests include the development of data analysis and data mining technology in ICT industry.

Masato Koda received the BS, MS, and PhD in Aeronautics and Astronautics all from the University of Tokyo. He was on the faculty of California Institute of Technology, and an Associate Professor at the University of Tokyo. He served as Professor of Systems and Information Engineering, University of Tsukuba, where he is now an Emeritus Professor.