Transactions on Machine Learning
and Data Mining
Vol. 5, No. 1 (2012) 3-22

ISSN:1865-6781,
ISBN: 978-3-942952-11-8

ibai Publishing
www.ibai-publishing.org

Contrasting Correlations by an Efficient Double-Clique Condition

Aixiang Li, Makoto Haraguchi, and Yoshiaki Okubo

Graduate School of Information Science and Technology
Hokkaido University
N-14 W-9, Sapporo 060-0814, JAPAN
{aixiang,makoto,yoshiaki}@kb.ist.hokudai.ac.jp

Abstract. Contrast set mining has been extensively studied to detect changes between several contrasted databases. Previous studies mainly compared the supports of an itemset and extracted the itemsets with significantly different supports across those databases. Differently, we contrast the correlations of an itemset between two contrasted databases and attempt to detect potential changes. Any highly correlated itemset is not of our concern in order to focus on implicitly emerging correlation. Therefore, we set correlation constraints (upper bounds) in both databases, and then extract the itemsets consisting of items that are not highly correlated in both databases, but exhibiting a potential change of correlations from one database to the other. We investigate both positive and negative correlations. We also investigate the correlation under conditioning by third variables. Thus, we also study so called partial correlation. To measure this kind of correlation, we use extended mutual information. In our search procedure for the correlated itemsets, we use a double-clique condition, which is necessary for itemsets to be solutions satisfying the correlation constraints. We show its usefulness through experiments.

Keywords: Contrast correlations, Extended mutual information, Double-Clique

1 Introduction

This paper[1] concerns a problem of mining contrast sets. Given several databases to be contrasted, detecting changes over these databases has been paid much attention recently. Pioneering studies [2–5] mainly compared the supports of an itemset and extracted the contrast sets with significantly different supports across these databases. However, there would still exist various changes that can not be detected by only contrasting supports. In this paper, we focus on another aspect of an itemset, *correlation of an itemset*, and attempt to detect valuable information obtained from correlation change. Particularly, we present a simple double-clique algorithm that enumerates itemsets that show higher but not too much higher correlation in one database and lower correlation in another for contrasting. Thus, several previous studies, *correlation mining* [6–11], *correlation change mining* [12], *emerging pattern mining* [4, 5] and *contrast set mining* [2, 3, 13] are related to our research.

The study of correlation mining began with research on basket data analysis [6, 7]. The task is to find minimal statistically correlated itemsets. The degree of correlation is measured via χ^2 statistics. "Minimally correlated sets" are defined to be non-redundant itemsets for which null hypotheses of independence are rejected based on the judgement from χ^2 value, where each item is regarded as a Boolean random variable.

The computation of χ^2 statistics for itemsets requires making and recording partitions of database object sets. The same problem similarly occurs with *k-way information* [14], an extended mutual information for more than two variables, which we also use to calculate the degree of our correlation. It is also used for *subspace clustering* [15]. The advantage of using χ^2 and k-way information is monotonicity. The degree of correlation increases as we add more items. Therefore, we can design a simple miner based on monotonicity. The disadvantage comes from the simple fact that we need to compute partitions of database object sets (refer to [16] Sec. 15.4). This requires a large amount of space and time. For pattern mining, we have to examine a large number of item combinations for which we have to make partitions. Due to this disadvantage, there does not seem to be many studies on pattern mining based on k-way information or χ^2 in spite of the property of monotonicity.

According to the literature [8–10, 17], another correlation measure for itemsets, which we call "bond" [17], is often used. The bond for itemsets increases as the intersection of extents of member items increases and their union decreases. Thus, the bond measure calculates the overlapping degree. To calculate the bond, it suffices to maintain the union and intersection of extents of member items. Therefore, the computation of the bond is much easier than that of χ^2 or k-way information. In addition, the bond satisfies the anti-monotonicity property that it decreases as we add more items. As a result, we can design effective pattern

[1] This paper is an extended version of [1]. In this paper, we describe the proposal for mining correlation contrast sets in further detail and provide more experimental results.

miners as in those literature [8–10].

In this paper, we use k-way information for solving correlation change problem over two or more databases, not for detecting correlated itemsets in a single database. The reason we do not use χ^2 is clear. As stated in [6, 7], χ^2 is only a statistical approximation for judgement. Comparing χ^2 values over different databases is not meaningful. We have two contrasted databases and prefer a measure that can be compared over different databases. Therefore, we use k-way information instead of χ^2.

The reasons we still use k-way information as a correlation measure is more essential. The first one is concerned with the interestingness of itemsets, and the other comes from an issue about pruning technique.

We first explain why k-way information measure is useful even for correlation change. Suppose we have a database of Japanese newspaper articles just after a catastrophic event, e.g., the 1995 Kobe Earthquake in Japan. According to the experiment described in Section 6, the mutual information ($k = 2$ way information) of "doctor" and "airport" is low. This implies that neither positive nor negative correlation between "doctor" and "airport" was observed in the articles even after the earthquake. It should be noted that each item is regarded as a Boolean random variable and the information measure will give a higher positive value for both the positive and negative correlations. On the other hand, under the conditioning of the term "Japan/Japanese", the correlation between "doctor" and "airport" increases to some extent. The documents containing these terms are about the fact that non-Japanese doctors flew into airports to perform medical service for the earthquake stricken Kobe area. To make the existence of doctors coming from countries outside of Japan explicit, the term "Japan/Japanese" was used negatively in the articles. Thus, the actual correlation of "doctor" and "airport" under the condition specified by "Japan/Japanese" is the correlation of "foreign doctor" and "airport". The conditioning given by "Japan/Japanese" makes such an unusual but meaningful correlation of words explicit. The correlation among these terms before the earthquake is very low. Therefore, we regard the itemset as one with "emerging correlation".

On the other hand, the bond for "doctor", "airport" and "Japan/Japanese" is low before and after the earthquake. There is no correlation change in the sense of bond. The reason is simple. The document set with the term "doctor" or "airport" is larger. In fact, there are relatively many articles about Japanese doctors for which we do not find "Japan/Japanese" that is a default term for Japanese doctors and is not explicitly written. In other words, from the correlation change, we can detect unusual but meaningful term relationships before and after an event that are difficult to detect by contrasting correlation in the sense of co-occurrence as the bond.

We summarize the above observations and show our basic specification as follows:

⟨**1**⟩ The target itemset X must have moderate correlation in a designated database DB_2 compared with its weak correlation in another database DB_1

that is to be contrasted. For this requirement, we use two parameters, δ_1 and δ_2 to make correlation constraints for itemset X and set a minimum correlation increase d, that $I_1(X) \leq \delta_1$, $I_2(X) \leq \delta_2$, and $I_2(X) - I_1(X) \geq d$, $\delta_2 > \delta_1 + d$, where $I_j(X)$ is the k-way information of X in DB_j, where k is the length of itemset X.

⟨**2**⟩ As an optional constraint, we use an extended support constraint based on that in [6, 7]. This is simply to exclude the itemsets that involve more rare (or common) items.

⟨**3**⟩ All the itemsets satisfying the above constraints are generated based on the double-clique pruning technique. After enumeration, only the itemsets X whose $I_2(X)$ is greater than $I_1(X)$ to some extent are outputted. Thus, the procedure presented in this paper is simply a generate-and-test method. The effectiveness of the double-clique pruning is therefore the key issue.

Double-clique pruning prevents from trying to add useless items to the present itemsets during the search process. From the monotonicity of k-way information, the mutual information of any paired items must be less than δ_1 in DB_1 and δ_2 in DB_2. This is a necessary condition for any pair of member items in solution itemsets. Consequently, solution itemsets are understood as cliques under both of δ_1 and δ_2. Under the first constraint by δ_1 (P1), we can exclude almost correlated combinations of items that do not depend on the event occurring just after DB_1 and just before DB_2. Under the constraint by δ_2 (P2), the characteristic and visible correlation in DB_2 will be cut off. Since the event has the power to change the correlational relationships among terms (items), the constraint (P2) becomes more effective. It should also be noted that any itemset moderately-correlated in both DB_1 and DB_2 can never form a double-clique because it would satisfy (P2) but not (P1). In other words, itemsets with moderate-correlations that are little affected by the event can be excluded by our double-clique pruning. If there are many such itemsets *almost unchanged*, the pruning can be more effective. We can fortunately observe a large number of such itemsets even for catastrophic events, such as the Kobe Earthquake, as presented in our experimentations.

Clearly the correlation in [6, 7] and ours focus on even negative events meaning that some items are anti-co-occurred. This point distinguishes ours research from emerging pattern mining [5] and contrast set mining [3, 13], all of which are mainly concerned with support change, not correlation change in the sense of this paper. Similarly, correlation change in the sense of that in [12] also differs from ours. In the case of [12], paired itemsets with larger *lift* are preferred. Needless to say, the *lift* is a kind of self-mutual information about positive events, so it is different from our notion of correlation. In addition to this definitional difference, they often find itemsets with very small support for which higher lift can be observed. We do not prefer such itemsets as they might be accidental and are consequently not of our concern. Furthermore, as we described above, our correlation is not limited to a pairwise one and is extended to that among three or more items. Therefore, our correlation study is also different from previous pairwise correlation studies (i.e., [11, 18]).

The remainder of this paper is organized as follows. In the next section, we

begin with the preliminary definitions and notations. Our correlation measure, k-way mutual information, and its property are presented in Section 3. In Section 4, we define our problem of mining correlation contrast sets. We describe our efficient double-clique search algorithm for extracting correlation contrast sets in Section 5. We show our experimental results in Section 6. Finally, we make concluding remarks in Section 7.

2 Preliminaries

Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of *items*. An *itemset* X is a subset of $\mathcal{I}$. If $|X| = k$, then X is called a k-itemset. A *transaction* T is a set of items, $T \subseteq \mathcal{I}$, and a *transaction database* $\mathcal{D}$ is a collection of *transaction*.

In this paper, we take an alternative view of transaction data in order to consider probabilities of itemsets. We regard an item as a Boolean random variable. Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of n *Boolean random variables*. A transaction T is an n-tuple in $\{0, 1\}^n$, where 1 represents *presence* of the corresponding item and 0 *absence* of the corresponding item in T. Thus, a database $\mathcal{D}$ is a collection of n-tuples in $\{0, 1\}^n$.

Based on the above definitions, the notion of supports of itemsets also changes from the standard definition in association rule mining [19]. Similar to the cell values in a *contingency table* in [6], for a k-itemset, there are 2^k cell values in its contingency table. These cell values are considered as supports of the itemset, i. e., a k-itemset has 2^k supports. We take the values of supports in percent as an estimation of probabilities to calculate our correlations defined in the next section.

More specifically, let $\mathcal{D}$ be a database. For an item (1-itemset) a, the number of transactions in $\mathcal{D}$, in which a is present, is denoted as $T(a)$, and the number of transactions, in which a is absent, is denoted as $T(\overline{a})$. The probability of a, denoted as $p(a) = p(a = 1)$, is estimated as $p(a) = T(a)/|\mathcal{D}|$. Similarly, we define $p(\overline{a})$ as $p(\overline{a}) = p(a = 0) = T(\overline{a})/|\mathcal{D}| = 1 - p(a)$. Thus, the database $\mathcal{D}$ is partitioned into 2 cells (partitions) by a with the probabilities $p(a)$ and $p(\overline{a})$.

For a 2-itemset $\{a, b\}$, we have 4 probabilities, $p(ab)$, $p(a\overline{b})$, $p(\overline{a}b)$ and $p(\overline{a}\overline{b})$. The probability $p(ab)$ is estimated as $T(ab)/|\mathcal{D}|$, where $T(ab)$ is the number of transactions in $\mathcal{D}$ in which both a and b are present. The probabilities $p(a\overline{b})$, $p(\overline{a}b)$ and $p(\overline{a}\overline{b})$ are estimated in the same way. The definitions can be extended for a k-itemset such that $k \geq 3$. Generally, for a k-itemset, $\mathcal{D}$ is partitioned into 2^k cells corresponding to the 2^k cell values in its contingency table. Based on the contingency table of an itemset, we calculate the correlation of an itemset measured via k-way mutual information.

3 Correlation Based on *k*-Way Mutual Information

To reveal the relationships between itemsets that cannot be expressed by association rule mining [19], *correlation mining* has been developed, and correlation has been measured via χ^2 value [6, 7], *lift* [12], $all\_confidence$, *bond* [17], NMI [11]

and *Pearson's correlation coefficient* [18]. In this paper, we consider the correlations between two or more items that are regarded as *Boolean random variables*. As the reason discussed in [7], we also do not use the correlation coefficient as a measure for our correlation. Usually, when we analyze the correlation between a pair of items or itemsets given a transaction database, we can observe the following cases. For a pair of items a and b, if $p(ab) > p(a)p(b)$ holds, they are correlated *positively*. Oppositely, if $p(ab) < p(a)p(b)$, they are correlated *negatively*. These pairwise correlations can be measured via *lift* or other previous pairwise measures.

Furthermore, the degree of correlation between items or itemsets is heavily affected by other items or conditions. For example, for three items a, b and c, we assume a correlation between a and b, denoted as $cor(a;b)$, is very low. By a condition of c, however, a correlation $cor(a;b|c)$ might become larger. In some cases, we might observe $cor(a;b|c) \gg cor(a;b)$. We assume that there exists a *partial correlation* between a, b under c. It should be noted that the meaning of this *partial correlation* is different from that in statistics. In this paper, it means a correlation caused by additional factors. Thus, our notion of *extended correlation* covers positive, negative and this kind of partial correlation among two or more items. Even though the *bond* can measure some correlation between two or more items, and the χ^2 statistic value can be used as a judgement for this kind of correlation, because of their limitations discussed in Section 1, we introduce k-way mutual information as a measure of our extended correlation based on information theory.

We consider both the absence and presence of an item in a transaction by regarding it as a Boolean random variable. Then our correlation between a pair of items is measured via standard *mutual information*. That is, for two items a and b, the correlation between them is calculated as $I(a;b)$. For three or more items, the correlation among them is calculated by an extended mutual information, called *k-way mutual information*.

Definition 1. (Itemset Correlation)
Let $\mathcal{D}$ be a database and $X = \{x_1, \ldots, x_k\}$ a k-itemset. The correlation of X in $\mathcal{D}$, denoted by $cor_{\mathcal{D}}(X)$, is measured via *k-way mutual information*, $I(x_1; \cdots ; x_k)$, which is defined as

$$\begin{aligned} cor_{\mathcal{D}}(X) &= I(x_1; \cdots ; x_k) \\ &= \sum_{x_1=0,1} \cdots \sum_{x_k=0,1} p(x_1 x_2 \ldots x_k) \log_2 \frac{p(x_1 x_2 \cdots x_k)}{p(x_1)p(x_2)\cdots p(x_k)}, \end{aligned}$$

where x_i is an item regarded as a Boolean random variable. ∎

The value of the extended correlation measured via k-way mutual information is calculated based on the probabilities of cell values in the contingency table of an itemset. It is not affected by the size of databases and smaller cell values (affecting factors of χ^2 values). Therefore, we can compare the correlations of an itemset across different databases. Its computation time complexity is $O(2^k)$, which is similar to that of χ^2 value computation.

It is easily proved that the extended correlation has the following property:

Proposition 1. Let X be a k-itemset. For any X' such that $X' \supseteq X$, we have

$$cor_{\mathcal{D}}(X') \geq cor_{\mathcal{D}}(X). \quad \blacksquare$$

The monotonicity of the extended correlations based on k-way mutual information provides an efficient pruning mechanism for our algorithm for mining correlation contrast sets, as discussed in the following sections.

4 Problem of Mining Correlation Contrast Sets

In this section, we define the *problem of mining correlation contrast sets*.

Our goal is to detect a potential change from one database $\mathcal{D}_1$ to the other $\mathcal{D}_2$ by contrasting correlations of itemsets in $\mathcal{D}_1$ with those in $\mathcal{D}_2$.

In general, itemsets consisting of highly correlated items in any database are easy to detect and would never be interesting to users. Therefore, we try to detect implicit changes of correlations constrained at not higher range in both databases. In other words, we prefer itemsets consisting of weakly correlated items in one database but moderately correlated in the other.

More specifically, we extract contrast sets which show potential increase of correlations from one database to the other under correlation constraints. For each of the contrasted databases, we provide a correlation upper bound. One is set to a lower level and the other to a medium level estimated after investigation of the correlation distribution between all item pairs in the corresponding databases. Then contrast sets with potential increases of correlations are extracted.

Our problem of mining correlation contrast sets is now formalized as follows:

Definition 2. (Correlation Contrast Sets)
Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be a pair of databases to be contrasted [2], d a *minimum correlation difference*, and δ_1 and δ_2 the correlation upper bounds for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, where $\delta_2 > \delta_1 + d$.

The *problem of mining correlation contrast sets* is to find all itemsets X satisfying the following constraints:

- Weak-correlation in $\mathcal{D}_1$, $cor_{\mathcal{D}_1}(X) \leq \delta_1$,
- Moderate-correlation in $\mathcal{D}_2$, $cor_{\mathcal{D}_2}(X) \leq \delta_2$, and
- Correlation increase, $cor_{\mathcal{D}_2}(X) - cor_{\mathcal{D}_1}(X) \geq d$. $\blacksquare$

5 Extracting Correlation Contrast Sets by Double-Clique Method

In this section, we present an algorithm for extracting correlation contrast sets defined in the previous section.

[2] We assume $\cup_{T \in \mathcal{D}_1} T = \cup_{T \in \mathcal{D}_2} T$, that is, the set of items that appear in $\mathcal{D}_1$ is the same as that in $\mathcal{D}_2$.

To find the contrast sets for a given pair of databases, we can simply enumerate all possible itemsets with a set enumeration tree, compute their correlations in both databases, and then select the itemsets with a defined increase of correlations. As we know, however, such a naive method would be impractical since the number of possible itemsets is exponential in the number of items. Moreover, as is similar to the construction of a contingency table for an itemset in [6, 7], the calculation of extended mutual information is also expensive, especially for longer itemsets. It is, therefore, necessary to provide some pruning mechanisms to exclude useless itemsets that can never be our targets.

5.1 Excluding Useless Itemsets with Double-Cliques in Anti-Correlation Graphs

An itemset X to be extracted must satisfy the correlation constraints bounded by δ_1 and δ_2, i.e., $cor_{\mathcal{D}_1}(X) \leq \delta_1$ and $cor_{\mathcal{D}_2}(X) \leq \delta_2$. It should be noted that our correlation monotonically increases as an itemset is expanded to its supersets. This monotonicity implies that for an itemset X, if $cor_{\mathcal{D}_i}(X) \leq \delta_i$, then any pair of items in X, a and b, always satisfy $cor_{\mathcal{D}_i}(\{a, b\}) \leq \delta_i$ where $i \in \{1, 2\}$. That is, the inequality for any item pair in X can work as a *necessary condition* for our target contrast sets. Therefore, if there is a pair of items a and b in X such that $cor_{\mathcal{D}_i}(\{a, b\}) > \delta_i$, X can never be a candidate of our targets. Such an itemset is not in the scope of our search and should be excluded for efficient computation of our contrast sets. In other words, it is *sufficient* for our computation to examine only itemsets X such that for any $a, b \in X$, $cor_{\mathcal{D}_i}(\{a, b\}) \leq \delta_i$ holds for each $i \in \{1, 2\}$. An itemset satisfying the condition can be easily identified as a *double-clique* in a pair of undirected graphs corresponding to two contrasted databases.

More specifically, for each $i \in \{1, 2\}$, let $G_{\mathcal{D}_i} = (\mathcal{I}, E_i)$ be an undirected graph, where $\mathcal{I}$ is the set of items appearing in both contrasted databases and the set of edges E_i is defined as

$$E_i = \{(a, b) \mid a, b \in \mathcal{I} \wedge cor_{\mathcal{D}_i}(\{a, b\}) \leq \delta_i\}.$$

We call such a graph an *anti-correlation graph* for $\mathcal{D}_i$. It is easy to see that for a clique Q in $G_{\mathcal{D}_i}$, the correlation between any item pair in Q is less than or equal to δ_i. Therefore, if an itemset X forms a clique in both $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$, called a *double-clique*, X corresponds to a candidate to be examined for contrast sets in our search. Thus, we only need to enumerate and examine double-cliques in both $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$ by excluding obviously useless itemsets. In our actual computation, double-cliques of $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$ can be found by simply finding cliques in the graph $G = (\mathcal{I}, E_1 \cap E_2)$.

5.2 Enumerating Cliques in Undirected Graph

For a given undirected graph $G = (V, E)$, every clique in G can be enumerated *systematically*. A clique in G is a subset of V. Moreover, any subset of a clique

is also a clique. We can, therefore, enumerate all cliques in G along with the *set-enumeration tree* for V [20].

Let us assume a *total ordering* on $V = \{x_1, \ldots, x_n\}$ defined as $x_i \prec x_{i+1}$ $(1 \leq i \leq n-1)$. For a clique $Q \subseteq V$ in G, Q can be expanded into a larger clique by adding a certain vertex to Q. Such a vertex to be added is precisely defined with the notion of *extensible candidates*.

Definition 3. (Extensible Candidates for Clique)
Let $G = (V, E)$ be a graph and Q a clique in G. A vertex $x \in V$ adjacent to any vertex in Q is called an *extensible candidate* for Q. The set of extensible candidates is denoted as $cand(Q)$, i.e.,

$$cand(Q) = \{x \in V \mid \forall y \in Q, \quad (x, y) \in E\} = \bigcap_{y \in Q} N_G(y),$$
where $N_G(y)$ is the set of vertices adjacent to y in G. ∎

Since it is obvious from the definition that for any extensible candidate $x \in cand(Q)$, $Q \cup \{x\}$ always becomes a clique, we can easily generate a larger clique of Q by adding $x \in cand(Q)$ such that $tail(Q) \prec x$, where $tail(Q)$ is the last (maximum) element in Q under the ordering $\prec$. Starting with the initial $Q = \phi$ and the initial set of extensible candidates $cand(Q) = cand(\phi) = V$, we can enumerate all cliques in G by expanding Q with $cand(Q)$ sequentially in a depth-first manner [21]. Clearly, the computation time of $cand(Q)$ is directly proportional to $\sum |N_G(y)|$, where $|N_G(y)|$ is the degree of y in G. By the double correlation constraints, the degree is reduced significantly, therefore, the computation of cliques becomes much faster than set enumeration by naive method.

5.3 Additional Support Constraint

A correlation based on the judgement with χ^2 statistics has been investigated in [6, 7]. In the framework, a support constraint to exclude itemsets, which cannot cause any significant gain of correlations is available. Fortunately, a similar support constraint is also valid for our correlation measured via k-way mutual information.

As has been discussed previously, for a k-itemset $X = \{x_1, \ldots, x_k\}$, a database $\mathcal{D}$ is partitioned into 2^k cells corresponding to those cells in the contingency table of X for computing the correlation of X. These cells are divided into two groups, one consists of the cells with $x_k = 1$ and the other consists of the cells with $x_k = 0$. The former is referred to as 1-half and the latter as 0-half. We assume that almost all the probabilities of 1-half cells or 0-half are very low. In this case, the captured item x_k could possibly be a rare or common item. From the definition of our correlation measured via k-way mutual information, expanding X cannot cause any significant gain of information (correlation). Therefore, X can be excluded from further expansion. This pruning is formalized as follows.

Definition 4. (Support Constraint)
Let $X = \{x_1, \ldots, x_k\}$ be a k-itemset. Itemset X has support s (in percent) at $p\%$ level if at least $p\%$ of the cells in both 0-half and 1-half determined by the k-th item x_k in X have supports s in each of the contrasted databases. ■

If an itemset X does not have support s at $p\%$ level given p and a minimum support s, then we do not need to expand X.

5.4 Algorithm for Extracting Correlation Contrast Sets

Based on the clique enumeration explained above, we can design a depth-first double-clique search algorithm for finding correlation contrast sets.

Given a pair of databases, $\mathcal{D}_1$ and $\mathcal{D}_2$, to be contrasted and a pair of correlation upper bounds for each database, δ_1 and δ_2, respectively, we first construct the anti-correlation graphs $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$. We then sequentially enumerate the double-cliques of $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$ in a depth-first manner. For each double-clique X, we check whether both $cor_{\mathcal{D}_1}(X) \leq \delta_1$ and $cor_{\mathcal{D}_2}(X) \leq \delta_2$ actually hold. If the constraints are satisfied, then the correlation increase of X is measured. If X has a significant increase of correlations from $\mathcal{D}_1$ to $\mathcal{D}_2$, we output X as one of our target contrast sets and then continue to expand X with $cand(X)$. If X cannot satisfy any of the correlation constraints and the support constraint, we do not have to expand X. We can immediately backtrack to another double-clique. This procedure is recursively iterated until no double-clique remains to be examined.

Our algorithm for detecting correlation contrast sets is summarized in Figure 1.

6 Experimental Results

The proposed algorithm was implemented in JAVA and evaluated with two types of databases, the *Mainichi News Articles* database and *BankSearch*[22] web document database. We show some interesting contrast sets actually extracted and then report on the computational performance of our algorithm on a PC with Core2 Duo E8500 and 4GB main memory.

6.1 Contrasted Databases

Mainichi News Articles is a collection of articles of a Japanese newspaper "*Mainichi*", especially in 1994 and 1995. Since at the beginning of 1995, there happened "*Kobe Earthquake*" in Japan, we wanted to discover potential changes before and after the earthquake. The city of "*Kobe*" was one of the most damaged cities. From the articles, we selected those including the keyword "*Kobe*". After a morphological analysis, we extracted only nouns and removed too infrequent and frequent words. The remaining 406 words were regarded as items. To obtain contrasted databases, the articles were divided into those of 1994 and those of

Input:
A pair of databases $\mathcal{D}_1$ and $\mathcal{D}_2$ to be contrasted,
A pair of upper bounds of correlations δ_1 and δ_2,
Minimum increase of correlations d,
Minimum support s at level p%;
Output:
The set of correlation contrast sets, $\mathcal{CCS}$;

Procedure `MAIN`()
$\mathcal{I} \leftarrow$ the set of items that appear in both $\mathcal{D}_1$ and $\mathcal{D}_2$ with a pre-defined order $\prec$;
Construct anti-correlation graphs $G_{\mathcal{D}_1}$ and $G_{\mathcal{D}_2}$ based on δ_1 and δ_2, respectively;
Double-clique $C \leftarrow \phi$;
$Cand \leftarrow \mathcal{I}$;
$\mathcal{CCS} \leftarrow \phi$;
`EXPAND`(C, $Cand$);
return $\mathcal{CCS}$

Procedure `EXPAND`(C, $Cand$)
for each $x \in Cand$ such that $tail(C) \prec x$ **do**
  $NewC = C \cup \{x\}$;
  Calculate $2^{|NewC|}$ supports (probabilities) in $\mathcal{D}_1$ and $\mathcal{D}_2$;
  if the supports of $NewC$ have s at level p% in both $\mathcal{D}_1$ and $\mathcal{D}_2$ **then**
    if $Cor_{\mathcal{D}_1}(NewC) \leq \delta_1$ and $Cor_{\mathcal{D}_2}(NewC) \leq \delta_2$ **then**
      if $Cor_{\mathcal{D}_2}(NewC) - Cor_{\mathcal{D}_1}(NewC) \geq d$ **then**
        $\mathcal{CCS} = \mathcal{CCS} \cup \{NewC\}$;
      end if
      $NewCand = Cand \cap (N_{G_{\mathcal{D}_1}}(x) \cap N_{G_{\mathcal{D}_2}}(x))$;
      `EXPAND`($NewC$, $NewCand$);
    end if
  end if
end for

Fig. 1. Depth-first double-clique search algorithm for extracting correlation contrast sets

1995. The former consists of 2343 articles and is referred to as $\mathcal{D}_{1994}$. The latter, referred to as $\mathcal{D}_{1995}$, has 9331 articles.

BankSearch is a collection of web documents from 11-categories [22]. We selected two themes "*Banking and Finance*" and "*Sports*" from the collection to obtain a pair of databases to be contrasted. The former consisting of "Commercial Banks", "Building Societies", and "Insurance Agencies" categories is referred to as $\mathcal{D}_{Bank}$ and the latter consisting of "Soccer","Motor Sport", and "Sport" categories $\mathcal{D}_{Sports}$. Both themes consist of 3000 web documents. After standard preprocessing (stemming, removing stop-words, and removing too frequent and infrequent words), we extracted 585 words as items, and further 1217, 2215, 3008, and 4076 words for a scalability experiment.

6.2 Extracted Correlation Contrast Sets

We present examples of extracted contrast sets that show potential increases of correlations from $\mathcal{D}_{1994}$ to $\mathcal{D}_{1995}$.

Some extracted correlation contrast sets are listed in Table 1 under the parameter settings $\delta_1 = 0.001$, $\delta_2 = 0.003$, $s = 0.005$, $p = 0.25$, and $d = 0.001$.

Table 1. Examples of extracted correlation contrast sets for $\mathcal{D}_{1994}$ and $\mathcal{D}_{1995}$ and their correlation values

contrast sets	*cor* in $\mathcal{D}_{1994}$	*cor* in $\mathcal{D}_{1995}$	increase from $\mathcal{D}_{1994}$ to $\mathcal{D}_{1995}$
{*university, support, Osaka*}	0.00095379	0.00195461	0.00100083
{*doctor, airport, Japan*}	0.00053963	0.00224158	0.00170195
{*children, water, Kansai*}	0.00010262	0.00287586	0.00277324
{*secure, baseball*}	0.00088372	0.00281824	0.00193452
{*player, death*}	0.00003282	0.00156204	0.00152922
{*service, coach*}	0.00008051	0.00109476	0.00101425
{*enterpriser, anxiety*}	0.00000002	0.00299607	0.00299605
{*building, water*}	0.00000156	0.00293023	0.00292867

It should be noted that most of the extracted contrast sets are concerned with the earthquake in 1995.

The first three itemsets, *Osaka*, *Kansai*, and *Japan* are the names of the places where the earthquake struck. In each of the itemsets, the correlation among the three component items is very low in 1994, but potentially increases in 1995. We found that for each itemset, we observed that *partial correlations* increase in 1995 between the first two component items given the third item. That is, the conditional correlation between the first two items x_1 and x_2 given the third item x_3, $cor(x_1; x_2|x_3)$, is greater than $cor(x_1; x_2)$ without the condition by x_3 in 1995. For example, for the itemset {*doctor, airport, Japan*}, *cor*(*doctor;airport*) is 0.0002 *bit* actually in 1995. On the other hand, we observed $cor(doctor; airport|Japan) = 0.0006$ in 1995. The items *doctor* and *airport* were almost not correlated, but become some correlated under the condition by *Japan/Japanese* after the earthquake. The original news articles related to those terms revealed the fact that non-Japanese doctors flew into airports to provide medical support after the earthquake. That is, in the articles that include *doctor* and *airport*, but rarely include *Japan/Japanese*(occurs negatively), *doctor* and *airport* become more correlated. In fact, the condition given by *Japan/Japanese* makes a meaningful correlation between *foreign doctor* and *airport*.

In addition to these itemsets, we also extracted contrast sets consisting of *negatively correlated* items. For example, we found that there were very few news articles that include *baseball* and *secure*, particularly in 1995. We checked the *interest*[6], $p(secure, baseball)/p(secure)\,p(baseball) = 0.5$ in 1994 and even 0.1 in 1995. We found that the two items correlated *negatively* in both years. Their

negative correlation, however, increases some significantly from 1994 to 1995. The other two examples $\{player, death\}$ and $\{service, coach\}$ in the table also show increases of negative correlations.

The two itemsets at the bottom of Table 1 are examples of positively correlated itemsets, their items are almost independent in 1994, but become some significantly correlated (positively) in 1995. Thus, all of the itemsets in Table 1 reveal some changes before and after the earthquake.

Similarly, we also extracted the three kinds of correlation contrasted sets from web documents databases after comparing $\mathcal{D}_{Bank}$ with $\mathcal{D}_{Sports}$. Every contrasted database consists of 3000 documents with 585 words. Some examples of extracted correlation contrast sets are listed in Table 2 at the parameter settings $\delta_1 = 0.0008$, $\delta_2 = 0.003$, $s = 0.005$, $p = 0.25$ and $d = 0.001$.

Table 2. Examples of extracted correlation contrast sets for $\mathcal{D}_{Bank}$ and $\mathcal{D}_{Sports}$ and their correlation values

contrast sets	*cor* in $\mathcal{D}_{Bank}$	*cor* in $\mathcal{D}_{Sports}$	increase from $\mathcal{D}_{Bank}$ to $\mathcal{D}_{Sports}$
{*qualify, easily, agency*}	0.00076276	0.00296581	0.00220305
{*assure, interact, partner*}	0.00047690	0.00265653	0.00217963
{*loan, race*}	0.00028523	0.00241107	0.00212584
{*car, loan*}	0.00000019	0.00112682	0.00112663
{*thank, partner*}	0.00000002	0.00299607	0.00299605
{*win, property*}	0.00000156	0.00293023	0.00292867

The first two itemsets are the partially correlated patterns. In these two itemsets, we can see that the partial correlations between the first two component items given the third item significantly increase in the web documents on sports. These documents associated with the former itemset mainly report on the qualification of players in sports and their photo and advertisement agencies, and the documents associated with the latter are mainly about the sports club partners and their interactive service. The next two itemsets are the negatively correlated patterns and the last two itemsets are the positively correlated ones. All these itemsets with a correlation increase distinguish the two different themes *Banking and Finance* and *Sport* and reveal the change from one theme to the other.

We should emphasize that about 30% of our extracted contrast sets whose supports (in usual concept) or bonds [17] change little or even decrease. This means that our contrast sets cover many itemsets that can never be extracted using support-based or bond-based methods. This is a major advantage of our method. Some examples are listed in Table 3. The original news articles related to these examples report that some professors were very active in the reconstruction of the Kansai region (where Kobe is located), many contractors participated in the reconstruction of Kansai and people from many places provided economic

Table 3. Correlation contrast sets and their bond and support values

contrast-sets	$bond_{1994}/bond_{1995}$	sup_{1994}/sup_{1995}	cor_{1994}/cor_{1995}
{active,professor,Kansai }	0.00365/0.00146	0.00085/0.00021	0.00036/0.00268
{contractor,participate,Kansai }	0.00131/0.00086	0.00085/0.00064	0.00024/0.00256
{aid,news,Osaka }	0.00191/0.00115	0.00042/0.00021	0.00004/0.00235

or other aid to Osaka (in Kansai region). All these itemsets reveal important information after the earthquake.

6.3 Computational Performance

To evaluate the computational performance of our algorithm, we compared it with a naive method that simply enumerates all the itemsets and then checks correlation values of the itemsets in both databases.

For the contrasted databases $\mathcal{D}_{1994}$ and $\mathcal{D}_{1995}$, we tested the efficiency of our algorithm at 10-pairs of correlation upper bounds $(\delta_1; \delta_2)$ in increasing order, where δ_1 was set to a lower level of correlations and δ_1 a medium level. These concrete values were determined based on the investigation of correlations between every pair of items.

Computation time for each $(\delta_1; \delta_2)$ at the parameter settings $s = 0.005$, $p = 0.25$ and $d = 0.001$ is shown in Figure 2. As the upper bounds of correlations becomes some higher, e.g., $(0.001; 0.003)$, the performance curve by the naive method tends to increase exponentially, while the curve by our algorithm shows a gradual increase.

The numbers of examined itemsets are also shown in Figure 3. By double-clique condition pruning, many almost correlated itemsets that do not depend on the event, moderately correlated itemsets that are little affected by the event, furthermore, characteristic and visible correlation in DB_2 are excluded. Therefore, the number of examined itemsets by our algorithm are at least 4 times fewer than the number of those by the naive method. These experimental results on news data show that our double-clique method is effective and well outperforms the naive method.

We also tested the efficiency of our algorithm on the contrasted databases $\mathcal{D}_{Bank}$ and $\mathcal{D}_{Sport}$ with 3000 documents respectively and 585 terms. Figure 4 and Figure 5 respectively compare the computation times and the numbers of examined itemsets with the naive enumeration method at the parameter settings $s = 0.005$, $p = 0.25$, $d = 0.001$ and 10-pairs of $(\delta_1; \delta_2)$. The web data are denser and have more items than the news data. However, we find that our double-clique search algorithm performs much better than the naive enumeration algorithm.

During double-clique search, the support constraint has been used as another pruning rule. To check the effectiveness of this pruning rule, We ran our algorithm and the naive algorithm with/without the support constraint, using the

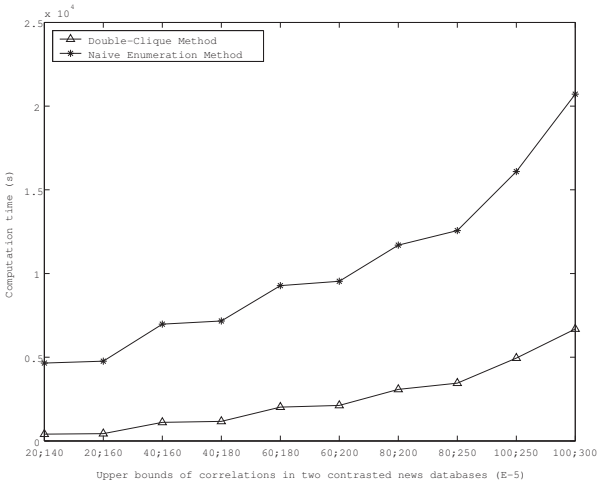

Fig. 2. Computation times on news data

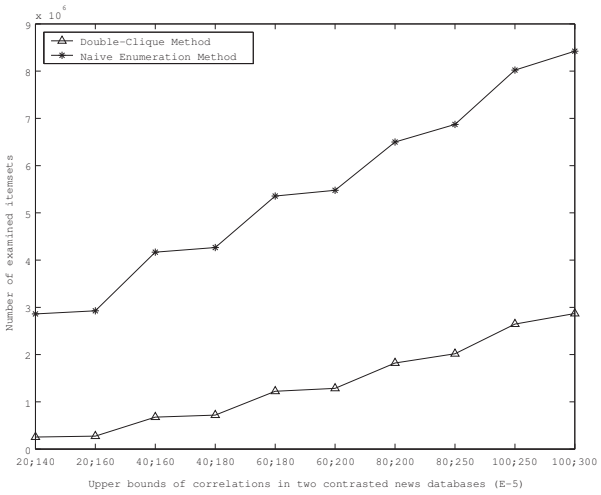

Fig. 3. Numbers of examined itemsets on news data

contrasted databases $\mathcal{D}_{Bank}$ and $\mathcal{D}_{Sports}$, at several $(\delta_1; \delta_2)$ settings and other parameter settings $s = 0.01$, $p = 0.25$, and $d = 0.0001$. We observed the effect of the support constraint. The numbers of examined itemsets by each algorithm are listed in Table 4.

Under each $(\delta_1; \delta_2)$ setting with/without the support constraint, the numbers of examined itemsets by our algorithm are just around $1.2 - 1.8\%$ of those by the

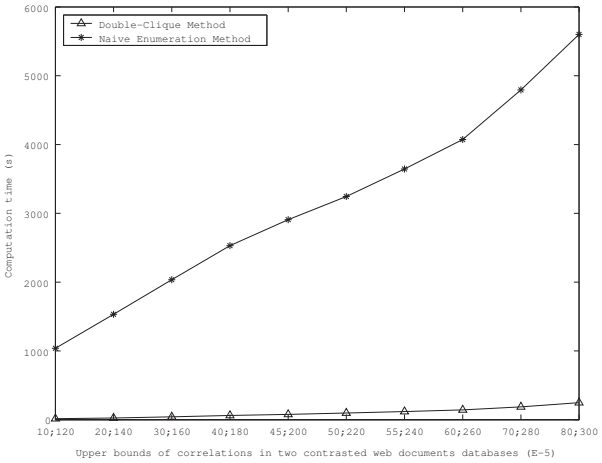

Fig. 4. Computation times on web data

naive method since many itemsets with unchanged correlations or too dramatically changed correlations were excluded by double-clique condition pruning. Furthermore, in both methods, the support constraint has little effect on pruning useless candidates under the given correlation upper bounds, which are set to a lower or medium level. If the upper bounds of correlations are set higher, it is expected that the support constraint will be more effective in reducing the number of itemsets to be examined. However, we are mainly interested in correlation changes in a relatively lower range of correlations. In this sense, the correlation constraints and further double-cliques condition are *primary* and the support constraint *secondary* in our framework.

Table 4. Number of examined itemsets for $\mathcal{D}_{Bank}$ and $\mathcal{D}_{Sports}$

δ_1; δ_2	Naive Method		Our Method	
	without sup. cons.	with sup. cons.	without sup. cons.	with sup. cons.
0.00001; 0.00015	1,894,398	1,894,398	22,342	22,342
0.00005; 0.00025	4,233,078	4232,673	54,113	54,111
0.00010; 0.00030	6,324,954	6,321,595	95,034	95,021
0.00012; 0.00035	7,532,777	7,523,278	121,294	121,217
0.00015; 0.00040	9,129,877	9,087,954	163,412	163,099

To evaluate the scalability of our algorithm, we prepared contrasted databases with 1000, 2000, and 3000 web documents of the themes “Banking and Finance”

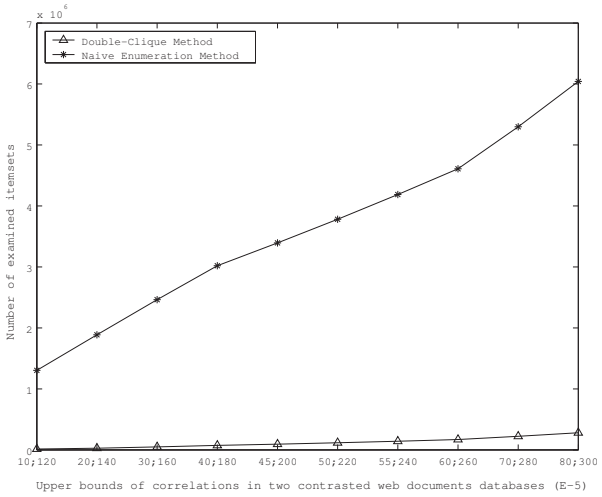

Fig. 5. Numbers of examined itemsets on web data

and "Sports" respectively from *BankSearch*. For each number of web pages, 585, 1217, 2215, 3008, and 4076 words were selected as items. We, therefore, had 15 pairs of contrasted databases with different scales. Then the databases with the same scales in both themes were contrasted.

We recorded the computation times for each of the 15-pairs of contrasted databases with three sets of correlation bounds $(\delta_1; \delta_2)$ at the parameter settings $p = 0.25$, $s = 0.001$, and $d = 0.0001$. The computation times are summarized in Figure 6.

The higher the correlation upper bounds become, the more computation times increase. Since a higher upper bound makes the corresponding anti-correlation graph denser, we are required to examine more itemsets. As mentioned above, however, we are not interested in the itemsets whose items are correlated at a higher level.

The number of words (items), n, also strongly affects computation times since the number of possible itemsets is exponential in n. Although our double-clique method is very effective, we might need to investigate more powerful pruning rules which can work better even for a much larger n. In contrast, the number of web documents (transactions) does not considerably affect computation time. From the experimental results in Figure 6, we can conclude that our algorithm can be considered scalable.

7 Conclusion and Further Research

In this paper, we give a proposal for contrasting correlations. We focused on implicitly correlated items in both of contrasted databases. We extracted the

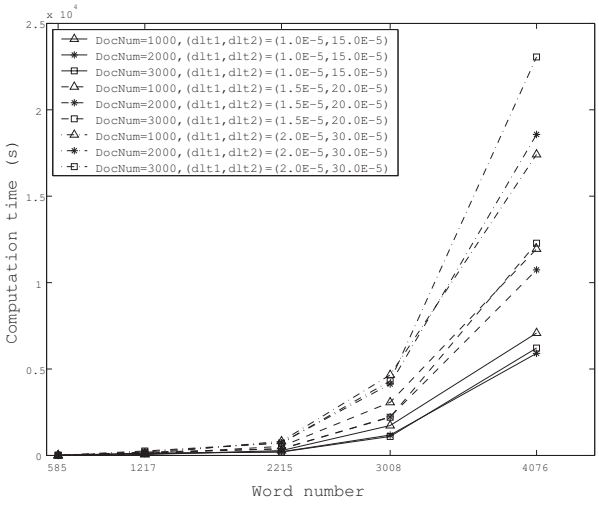

Fig. 6. Computation times by our algorithm on web data

patterns (itemsets) consisting of items correlated at a low level in one database but correlated at a medium level in the other database. The difference of the correlations between extracted items reveal the potential change that we attempt to make clear by comparing two (or more) databases. Differently from the previous contrast set mining and correlation studies, we contrasted the correlation among a set of items, and our information-theoretic correlation covers positive, negative and further partial correlation that actually exist among two or more items. To measure the extended correlation, we proposed an alternative measure - k-way mutual information that can be compared over different databases. To focus on implicit correlation changes, we assumed upper bounds of correlations in two contrasted databases and a minimum correlation increase. We designed a depth-first double-clique search algorithm for finding those contrast sets with potential correlation change. By this algorithm, useless itemsets can be excluded based on the double-clique pruning. In our experimentations, we extracted the contrast sets that cannot be found by contrasting supports or other correlation measures. However, these patterns are expected to bring valuable and interesting information when we try to compare two (or more) databases. Furthermore, we verified that our double-clique condition pruning is very effective, especially for excluding a large number of useless itemsets with moderate correlations little affected even by a big event.

In our current framework, we enumerated all contrast sets satisfying the correlation constraints. Although the double-clique pruning is quite effective, it is easily expected that the number of target itemsets will be quite large when we are given larger scale databases with more items. Additional constraints may be required to improve the quality of our targets that maintain reasonable mean-

ingfulness. Moreover, the top-N approach [23] may be promising. For example, our problem can be reformulated as a top-N optimization problem in which we try to minimize correlations in DB_1 keeping moderate correlations in DB_2 as a constraint so that we can extract itemsets with larger correlation changes. By investigating in these directions, our proposal for detecting correlation contrast sets would be further improved and become more useful.

Acknowledgments

This research was partially supported by a grant for the Hokkaido University Global COE program, "Next-Generation Information Technology Based on Knowledge Discovery and Knowledge Federation", from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Li, A., Haraguchi, M., Okubo, Y.: Contrasting Correlations by an Efficient Double-Clique Condition In: P. Perner (Ed.) MLDM 2011, LNAI, Vol. 6871, pp. 469-483, Springer-Verlag, Berlin Heidelberg (2011)
2. Bay, S.D., Pazzani, M.J.: Detecting Change in Categorical Data: Mining Contrast Sets. In: Proceedings of KDD'99, pp. 302-306, ACM Press, New York (1999)
3. Bay, S.D., Pazzani, M.J.: Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery 5, 213-246 (2001)
4. Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: Proceedings of KDD'99, pp. 43-52, ACM Press, New York (1999)
5. Dong, G., Li, J.: Mining Border Descriptions of Emerging Patterns from Dataset Pairs. Knowledge and Information Systems 8(2), 178-202 (2005)
6. Brin, S., Motwani,R., Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations. In: ACM SIGMOD International Conference on Management of Data, pp. 265-276, ACM Press, New York (1997)
7. Silverstein, C., Brin, S., Motwani,R.: Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. Data Mining and Knowledge Discovery 2, 39-68(1998)
8. Younes, N.B., Hamrouni, T., Yahia, S.B.: Bridging Conjunctive and Disjunctive Search Spaces for Mining a New Concise and Exact Representation of Correlated Patterns. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, Vol. 6332, pp. 189-204, Springer, Heidelberg (2010)
9. Kim, W.Y., Lee, Y.K., Han, J.W.: CCMine: Efficient Mining of Confidence-Closed Correlated Patterns. In: Dai, H., Srikant, R.,Zhang, C.(eds.) PAKDD 2004. LNCS(LNAI), Vol. 3056, pp. 569-579, Springer, Heidelberg (2004)
10. Zhu, F., Yan, X., Han, J., Yu, P.S., Cheng, H.: Mining Colossal Frequent Patterns by Core Pattern Fusion. In: 23rd IEEE International Conference on Data Engineering, pp. 706-715, IEEE Press, Los Alamitos (2007)
11. Ke, Y.P., Cheng, J., NG, W.: Mining Quantitative Correlated Patterns Using an Information-Theoretic Approach. In: ACM KDD'06, pp. 227-236 (2006)
12. Taniguchi, T.: A Study on Correlation Mining Based on Contrast Sets. Doctoral Dissertation, IST, Hokkaido University, Japan (2008)

13. Novak, P.K., Lavrac, N., Webb, G.I.: Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research 10, 377-403 (2009)
14. Zhang, X., Pan, F., Wang, W., Nobel, A.: Mining Non-Redundant High Order Correlations in Binary Data. In: Proceedings of VLDB, Vol.1(1), pp.1178-1188 (2008)
15. Cheng C., Fu A., Zhang,Y.: Entropy-Based Subspace Clustering for Mining Numerical Data. In: 5th ACM SIGKDD, pp. 84-93, ACM press, New York (1999)
16. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadephia (2007)
17. Omiecinski,E.: Alternative Interest Measures for Mining Associations in Databases. In: IEEE Transactions on Knowledge and Data Engineering 15, 57-69 (2003)
18. Ke, Y.P., Cheng, J., NG, W.: Correlation Search in Graph Databases. In: ACM KDD'07, pp. 390-399 (2007)
19. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD in 1993, pp. 207-216 (1993)
20. Rymon, R.: Search through Systematic Set Enumeration. In: International Conference on Principles of Knowledge Representation Reasoning. pp. 539-550, Morgan Kaufmann Publisher, CA (1992)
21. Tomita, E., Seki, T.: An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique with Computational Experiments. Journal of Global Optimization 37(1), 95-111(2007)
22. Sinka, M.P., Corne, D.W.: A Large Benchmark Dataset for Web Document Clustering. Soft Computing Systems: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications 87, 881-890 (2002)
23. Haraguchi, M., Okubo, Y.: Pinpoint Clustering of Web Pages and Mining Implicit Crossover Concepts. In: Zeeshan-ul-hassan Usmani (ed.) Web Intelligence and Intelligent Agents, pp. 391-410, InTech (2010)