Nr.: FIN-007-2008

A Temporal Extension of Closed Item Sets for Change Mining

Mirko Böttcher, Martin Spott, Rudolf Kruse

Arbeitsgruppe Computational Intelligence (IWS)

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

Technical Report

Impressum (§ 10 MDStV):

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Der Dekan

Verantwortlich für diese Ausgabe:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Prof. Dr. Rudolf Kruse
Postfach 4120
39016 Magdeburg
E-Mail: kruse@iws.cs.uni-magdeburg.de

http://www.cs.uni-magdeburg.de/Preprints.html

Auflage: 71

Redaktionsschluss: August 2008

Herstellung: Dezernat Allgemeine Angelegenheiten,
Sachgebiet Reproduktion

Bezug: Universitätsbibliothek/Hochschulschriften- und
Tauschstelle

A Temporal Extension of Closed Itemsets for Change Mining

Mirko Böttcher
University of Magdeburg
Faculty of Computer Science
39106 Magdeburg, Germany
miboettc@iws.cs.uni-magdeburg.de

Martin Spott
BT Group
Intelligent Systems Research Centre
Adastral Park
Ipswich, IP5 3RE, UK
martin.spott@bt.com

Rudolf Kruse
University of Magdeburg
Faculty of Computer Science
39106 Magdeburg, Germany
kruse@iws.cs.uni-magdeburg.de

Abstract

Frequent pattern mining often produces a vast set of results. To overcome this problem, two fundamental approaches are commonly employed: condensed representations, such as closed itemsets, and relevance assessment. In recent years, the change of itemsets over time is gaining increasing attention as a promising basis for developing novel, more comprehensible relevance assessment methods. One of the unsolved problems is that typically many of the observed changes are the side-effect of other changes. Existing condensed representation approaches fail in removing such redundancies because they have not been developed with the temporal dimension in mind. This paper proposes a novel approach for a condensed representation of itemsets which is based on utilizing temporal redundancies. In particular we prove that our approach yields a temporally non-redundant subset of closed itemsets which we therefore call temporally closed itemsets. Our experiments with real-life data sets show that the set of temporally closed itemsets is significantly smaller than the set of closed itemsets.

1. Introduction

Frequent pattern mining originally has been developed for market basket analysis, where each basket, also referred to as a transaction, consists of a set of purchased items [1]. Here, the goal of frequent pattern mining is to detect sets of items which frequently occur together and, in a subsequent step, to form rules which predict their co-occurrence. However, frequent pattern discovery is not only bound to the specific purpose of association mining. It can be applied to every relational database and plays also an essential role in fields such as sequential patterns [3] and episode discovery from event sequences [10].

The comprehensibility and utility of *frequent itemsets*, as frequent patterns are also called due to their roots, contribute much to their popularity. It is also well known that the number of discovered itemsets can be vast and thus difficult to examine by a user. Moreover, many of the itemsets will be obvious, already known, or not relevant.

Two fundamental techniques have been proposed to tackle this problem. First, condensed representation algorithms aim to produce a reduced number of itemsets from which all other itemsets can be derived. The probably most well-known condensed representation are *closed itemsets* [14]. Second, a variety of methods for relevance assessment have been developed which aim at providing a ranking of the itemsets according to their (likely) relevance to a user (cf. [18]).

In recent years, there has been an increasing research interest in methods which rate the relevance of itemsets by analyzing their change over time. Such methods are based on time series (also called histories) of support [12, 5]. Itemsets which change hint at unknown or surprising changes in the underlying population. Such changes may indicate that an intervening action is required [9], for instance, to rectify a problem. On the other hand, an itemset which always remains stable can be expected to describe an invariant of the population. Invariants, however, are almost always known by domain experts and are thus of less interest. Nevertheless, this approach suffers from the problem that many of the observed changes are simply the side-effect of other changes. For this reason it is desirable to first obtain a condensed representation which captures the fundamental set of itemset histories and allows to reconstruct those properties of all other itemsets and their histories that are necessary for change analysis.

Existing reduction techniques, such as closed itemsets, are not the optimal choice when used in this setting because they cannot detect nor utilize redundancies which are only visible when itemsets are analysed over time. Hence, they do not allow to reduce the number of itemsets towards the maximum possible extent. Consider, as an example, survey data which contains information about used telecommunication services, like broadband or phone, and the social background of customers, like their gender. Itemset discovery is applied to this dataset to discover usage patterns in a sociographical context. Assume that the following itemsets have been discovered:

$$X_1 : \text{BROADBAND=YES}$$

$$X_2 : \text{BROADBAND=YES, PHONE=YES}$$

Closed itemset discovery would detect that the itemset X_1 is redundant (i.e. non-closed) if it has the same support as X_2. This redundancy is due to the fact that a supplier may *always* bundle a broadband with a phone connection. This, in turn, is an *invariant* of the underlying domain and thus probably known to a domain expert. Now consider the itemset

$$X_3 : \text{BROADBAND=YES, GENDER=MALE}$$

and assume that its history of support values shows an upward trend. Using closed itemsets X_1 would be regarded as non-redundant (i.e. closed) with respect to X_3 because broadband users are *not always* males. Nevertheless, the *fraction* of males among all people who use broadband may be invariant over time. This means, X_1 and X_3 show qualitatively the same trend which has its root in X_1. The history of X_3 could be derived from the one of X_1 by multiplying it with a gender-related constant factor. For this reason, one of the itemsets is *temporally redundant* with respect to the other.

In this paper we propose *temporally closed itemsets* as an approach which accounts for such temporal redundancies. It extends the idea of closed itemsets towards the temporal dimension. As the central theorem of this paper we prove that the set of temporally closed itemsets is a subset of the set of closed itemsets. It results from removing redundancies from the set of closed itemsets which are only visible when itemsets are analysed over time. Our approach results in a set of itemsets which is minimal in the sense that the shape of every other itemset's history can be reconstructed from it. This information, in turn, is sufficient for subsequent change analysis if we assume that an itemset's relevance is primarily determined by its change over time. An itemset is declared as interesting not by the extent with which it exceeds a user-defined support threshold but by the qualitative way in which it changes [9, 2]. We show experimentally that mining temporally closed itemsets can lead to a significantly smaller result set than mining for closed itemsets.

The remaining of this paper is organized as follows. In Section 2 we discuss related work. Section 3 and Section 4 introduce the necessary background on frequent itemset mining and closed itemsets. In Section 5 we define temporal redundancy by introducing the concept of *temporally derivable itemsets*, which we will subsequently use in Section 6 as basis for the definition of the set of *temporally closed itemsets*. Section 7 discusses a statistical test for temporal closedness. Section 8 shows the experimental results we obtained.

2. Related Work

The approach described in this paper is related to two so-far rather distinct fields of association mining: condensed representations and change mining. For this reason we will first provide an overview over existing condensed representation approaches, followed by the necessary background on change mining methods for associations.

As already pointed out earlier, the number of discovered itemsets is usually vast and thus often hardly manageable by a user. For this reason, several approaches have been proposed which lead to a condensed representation of the set of discovered itemsets by utilizing redundancies such that all other itemsets can be derived from the representation. Four such techniques can be found in the literature: closed itemsets [14, 15, 19], counting inference [4], deduction rules [8] and disjunction free sets [7]. From the perspective of analyzing the change of itemsets over time these methods treat each element of a sequence of temporally ordered data sets independently from each other. For this reason, they do not have the capability to detect redundancies which are only visible if itemsets are analysed over time. Of these condensed representation approaches, closed itemsets are related to our approach which yields a subset of them. We will discuss closed itemsets in more detail in Section 4.

Several methods have been proposed in the area of association mining which aim to discover interesting changes in histories of itemsets and association rules, respectively. Agrawal et al [2] proposed a query language for shapes of histories. Liu et al [12] showed how trend, semi-stable and stable rules can be distinguished using a statistical approach. In [9] the temporal description length of an itemset is introduced which rates support changes by using methods from information theory. Frameworks to monitor and analyse changes in support and confidence are described in [17, 5]. All of these publications have in common that they employ a concept of itemset interestingness which is only based on the qualitative change of an itemset over time but not on the extent with which this itemset exceeds a user-defined support-threshold. None of these publications discusses how the set of discovered itemsets can be effectively reduced such that the shape of all other itemsets can still be derived, nor do they discuss how existing reduction techniques for itemsets can be extended towards the temporal dimension.

In [11] a method to detect so-called *fundamental rule changes* is presented that aims to identify changes in support and confidence of association rules which cannot be explained by other changes. The authors provide heuristic criteria for solving this task. However, their approach differs to our approach of temporally closed itemsets in the following aspects: first, their approach can only be applied to histories of two periods length, whereas much longer histories are the norm when analyzing change. An extension to many periods is not straightforward due to the form of the underlying statistical test. Second, due to the heuristic nature of their approach it can lead to counter-intuitive results [6].

3. Itemsets and Support Histories

Formally, itemset discovery is applied to a data set of *transactions*. Every transaction T is a subset of a set of items L. A subset $X \subseteq L$ is called *itemset*. It is said that a transaction T *supports* an itemset X if $X \subseteq T$. If $X \subset Y$ holds for two itemsets X and Y we will say that X is *more general* than Y because X puts less restrictions on the underlying transaction set. Likewise, we say that Y is *more specific* than X. Furthermore, we define $XY := X \cup Y$ for simplicity.

The statistical significance of an itemset X is measured by its *support* $\mathrm{supp}(X)$ which estimates $P(X \subseteq T)$, or short $P(X)$. It is said that an itemset is *frequent* if its support is greater than or equal to a user-defined minimum support value $\mathrm{supp}_{\mathrm{min}}$. The *downward closure property* of itemsets states that for two itemsets $Y \supset X$ the support of X is greater or equal to the one of Y, i.e. $\mathrm{supp}(X) \leq \mathrm{supp}(Y)$.

The change of an itemset is defined by the change of its support over time. The time series of support values is called *support history*. Formally, let D be a time-stamped data set and $[t_0,\ t_n]$ the minimum time span that covers all its tuples. The interval $[t_0,\ t_n]$ is divided into $n > 1$ non-overlapping periods $T_i := [t_{i-1}, t_i]$, such that the corresponding subsets $D_i \subset D$ each have a size $|D_i| \gg 1$.

After carrying out frequent itemset discovery for each $D_i, i = 1, \ldots, n$ the support of each itemset X is now related to a specific time period T_i. We will indicate this by using the notation $\mathrm{supp}_i(X) \approx P(X|\ T_i)$. An itemset X which has been discovered in all periods is therefore described by n support values. Imposed by the order of time the values form sequences $(\mathrm{supp}_1(X), \ldots \mathrm{supp}_n(X))$. These *support histories* capture many of the changes of the underlying domain. Hence, they are mostly not stable but exhibit trends and other patterns.

4. Closed Itemsets

Closed itemsets are a subset of itemsets from which all other itemsets can be derived without further mining. The formal underpinnings of closed itemset algorithms can be found in the theory of lattices and Galois connection closures [14]. Still, their meaning is rather intuitive: a closed itemset is the largest itemset common to a set of transactions. All non-closed itemsets have the same support as their closure, which is the smallest closed itemset containing them. Formally, a closed itemset is defined as follows (cf. [15]):

Definition 1 (Closed Itemset) *An itemset X is a closed itemset iff there exists no proper superset $Y \supset X$ such that* $\mathrm{supp}(X) = \mathrm{supp}(Y)$.

Several algorithms have been proposed to efficiently discover the set of closed itemsets from a given data set, for example: A-Close [14], Closet [15] and Charm [19].

In the context of analyzing changes of itemsets, closed itemsets have several shortcomings. First of all, as already mentioned in Section 2, they only take into account each data set separately. In fact, they were developed to be applied only for single data sets. As a result they do not account for redundancies imposed by the temporal dimension as the example in the Introduction showed. Secondly, the definition of closed itemsets as well as the proposed mining algorithms rely on strict equality between support values which makes closed itemset mining susceptible to a low data quality. One bad record can turn an actually non-closed itemset into a closed one. Here, a less restrictive comparison is desirable, for instance, on the basis of statistical tests.

5. Temporally Derivable Itemsets

As laid out in the Introduction, the aim is to find a set of itemsets which is non-redundant in the sense that it is the minimal set necessary to derive the shape of the history of all remaining itemsets. We therefore first have to define what makes a history of an itemset XY derivable from the history of X and thus the itemset XY *temporally derivable*:

Definition 2 (Temporally Derivable Itemset) *Let $XY, X \neq \emptyset$ be an itemset with support history $(\operatorname{supp}_1(XY), \ldots, \operatorname{supp}_n(XY))$. The itemset XY is said to be temporally derivable with regard to an itemset X, denoted $X \hookrightarrow XY$, iff for each $XZ, Z \subseteq Y$ with support history $(\operatorname{supp}_1(XZ), \ldots, \operatorname{supp}_n(XZ))$ there exists a constant $\epsilon, 0 < \epsilon \leq 1$ such that $\operatorname{supp}_i(XY) = \epsilon \operatorname{supp}_i(XZ), i = 1, \ldots, n$.*

The main idea behind the definition is that the history of an itemset and hence the itemset itself is temporally derivable if it has the same shape as the history of a more general itemset apart from a scaling factor ϵ. To emphasize the scaling factor ϵ we will sometimes use the notation $X \overset{\epsilon}{\hookrightarrow} Y$. The criterion $\operatorname{supp}_i(XY) = \epsilon \operatorname{supp}_i(X), i = 1, \ldots, n$ used within the definition can be rewritten as $\epsilon = \operatorname{supp}_i(XY) / \operatorname{supp}_i(X) = P(XY \mid T_i)/P(X \mid T_i) = P(Y \mid XT_i)$. This means, the probability of Y is required to be constant over time given X, so the fraction of transactions containing Y additionally to X constantly grows in the same proportion as X. In other words, the confidence (represented by the scaling factor ϵ) of the rule $X \rightarrow Y$ does not change over time. Such time-invariant properties, however, often represent domain knowledge known to a user. Thus, a user would be able to infer the history of XY if he knows the one of X. In the opposite direction, he could also derive the history of X from the one of XY.

Figures 1 and 2 show an example of a temporally derivable itemset taken from the customer survey data used for our experiments, cf. Section 8. For reasons of data protection, the underlying itemset cannot be revealed. For illustration, the reader is referred to the example given in the Introduction, instead. Figure 1 shows the support histories of the less specific itemset at the top and the more specific itemset below, both over 20 time periods. The shape of the two histories is obviously very similar and it turns out that the history of the more specific itemset XY can approximately be determined using the more general one X by applying a scaling factor. As shown in Figure 2, the reconstruction is not exact. The reason for this is noise. As a result, a statistical test is employed in Section 7 to test for temporal derivability. Obviously, the history of the less specific itemset could be determined from the more specific in the same way. In the following we will show several properties of temporally derivable itemsets which we will use later on in this paper:

Lemma 1 *All itemsets are temporally derivable with regard to themselves, i.e. $X \hookrightarrow X$.*

Proof 1 *Lemma 1 follows directly from Definition 2.*

Lemma 2 *Let $X \overset{\epsilon_1}{\hookrightarrow} XY$ and $X \overset{\epsilon_2}{\hookrightarrow} XZ$ with $Y \subset Z$ then $\epsilon_2 \leq \epsilon_1 \leq 1$, i.e. ϵ_i are a monotonously decreasing series.*

Proof 2 *By Definition 2 it is $\epsilon_1 \operatorname{supp}_i(X) = \operatorname{supp}_i(XY)$ and $\epsilon_2 \operatorname{supp}_i(X) = \operatorname{supp}_i(XZ)$. Using the downward closure property of itemsets $\operatorname{supp}(XY) \geq \operatorname{supp}(XZ), XY \subseteq XZ$ it follows that $\epsilon_1 \operatorname{supp}_i(X) \geq \epsilon_2 \operatorname{supp}_i(X)$. This, in turn, yields $\epsilon_1 \geq \epsilon_2$. Next, we show that $\epsilon \leq 1$. By Definition 2 it is $\epsilon \operatorname{supp}_i(X) = \operatorname{supp}_i(XY)$. From this it follows that $\epsilon \operatorname{supp}_i(X) \leq \operatorname{supp}_i(X)$ using the downward closure property of itemsets $\operatorname{supp}_i(XY) \leq \operatorname{supp}_i(X)$. Division yields $\epsilon \leq 1$.*

■

Lemma 3 *If $X \overset{\epsilon_1}{\hookrightarrow} Y$ and $Y \overset{\epsilon_2}{\hookrightarrow} Z$ then $X \overset{\epsilon_1 \epsilon_2}{\hookrightarrow} Z$, i.e. derivability is transitive.*

Proof 3 *By Definition 2 it is $\epsilon_1 \operatorname{sup}_i(X) = \operatorname{sup}_i(Y)$ and $\epsilon_2 \operatorname{sup}_i(Y) = \operatorname{sup}_i(Z)$. Substitution yields $\epsilon_1 \epsilon_2 \operatorname{sup}_i(X) = \operatorname{sup}_i(Z)$ and thus $X \overset{\epsilon_1 \epsilon_2}{\hookrightarrow} Z$.*

6. Temporally Closed Itemsets

Building upon the notion of temporally derivable itemsets we can define the set of non-redundant itemsets. If we assume that the interestingness of an itemset is solely determined by the changes represented in its history both X and XY would have the same interestingness if XY is temporally derivable from X. For example, in Figure 1 both histories show all characteristic features that would make them interesting for a user: a trend turning point and a declining, respectively inclining, trend left and right from it. Hence, if one is known the other(s) can be regarded as redundant.

Commonly, sequences of itemsets $X_1 \hookrightarrow X_2 \ldots \hookrightarrow X_n$ temporally derivable from each other are discovered. Thereby, we assume that this sequence is maximal in the sense that there exists no $Y \subset X_1$ or $Z \supset X_n$ such that $Y \hookrightarrow X_1$ or $X_n \hookrightarrow Z$, respectively. From such a sequence we will define the maximum element X_n as being non-redundant and treat the others as redundant. We will call such non-redundant itemsets *temporally closed itemsets* because they are related to closed itemsets as we prove later in this section.

Definition 3 (Temporally Closed Itemset) *An itemset X is temporally closed iff there exists no itemset $Y \supset X$ such that $X \hookrightarrow Y$. A temporally closed itemset is* frequent *if it exceeds a user defined support threshold in all periods.*

Apparently, from the above sequence $X_1 \hookrightarrow X_2 \ldots \hookrightarrow X_n$ the minimum element X_1 could also have been chosen as the non-redundant element. Nevertheless, the choice of the maximum X_n as the basis for the definition of temporally closed itemsets provides the advantage that in this way they can be related to closed itemsets and thus extending this established notion by temporal considerations.

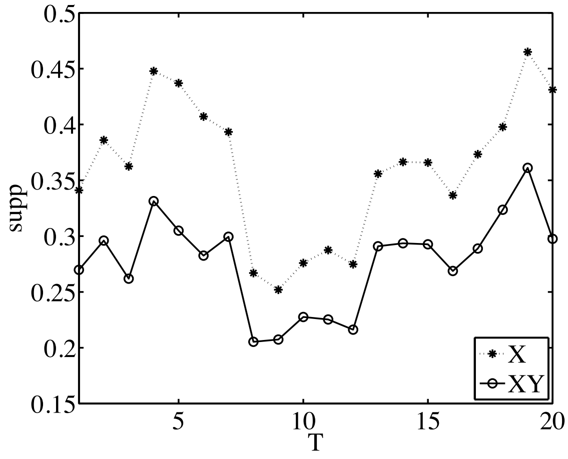

Figure 1. Histories of the itemsets XY and X showing that $X \hookrightarrow XY$

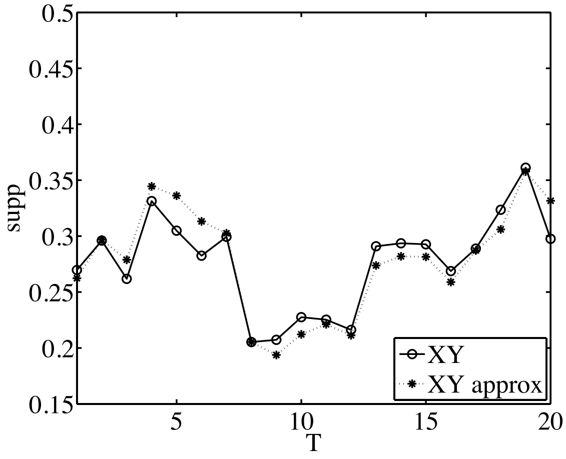

Figure 2. Approximated history of XY using the history of X

To show how the definition of temporally closed itemsets relates to closed itemsets discussed in Section 4 we will first extend Definition 1 such that it can be applied to histories of support.

Definition 4 (Closed over a Sequence of Time Periods) *An itemset X is closed over the sequence of time periods $\{T_1, \ldots, T_n\}$ iff there exists no itemset $Y \supset X$ such that $\mathrm{supp}_i(X) = \mathrm{supp}_i(Y), i = 1, \ldots, n$.*

In the following we will refer to itemsets which are closed over a sequence of time periods simply as closed itemsets.

By comparing Definition 4 with Definition 2 it can be seen that an itemset's closedness over a sequence of time periods can also be expressed using the notion of temporal derivable.

Lemma 4 *An itemset X is closed over the sequence of time periods $\{T_1, \ldots, T_n\}$ iff there exists no itemset $Y \supset X$ such that $X \overset{1}{\hookrightarrow} Y$.*

Proof 4 *Follows directly from the definition of a temporally derivable itemset (cf. Definition 2).*

We now have the necessary tools to prove the central theorem of this paper which shows that temporally closed itemsets are a subset of closed itemsets.

Theorem 1 *Let C be the set of all closed itemsets over the sequence of time periods $\{T_1, \ldots, T_n\}$ and TC be the set of temporally closed itemsets. Then, it is $TC \subseteq C$.*

Proof 5

$$\begin{aligned} X \in TC & \overset{Def.\ 3}{\Longleftrightarrow} & \nexists Y \supset X : \exists \epsilon \in (0,1] : X \overset{\epsilon}{\hookrightarrow} Y \\ & \Longrightarrow & \nexists Y \supset X : X \overset{1}{\hookrightarrow} Y \\ & \overset{Lemma\ 4}{\Longleftrightarrow} & X \in C \end{aligned}$$

From $X \in TC \Rightarrow X \in C$ it follows that $TC \subseteq C$.

The following counterexample shows that TC is generally a proper subset of C. Consider the itemsets $X_1 \subset X_2 \subset X_3 \subset X_4$ with $X_1, X_3 \in C$. Further, assume that $X_1 \overset{0.5}{\hookrightarrow} X_2 \overset{1}{\hookrightarrow} X_3 \overset{0.5}{\hookrightarrow} X_4$. Using Lemma 3 it is $X_1 \hookrightarrow X_4$ and $X_3 \hookrightarrow X_4$. Using Definition 3 it follows that $X_1 \notin TC$ and $X_3 \notin TC$.

This means, every temporally closed itemset is also a closed itemset but not every closed itemset is also a temporally closed one. The counterexample shows that one temporally closed itemset can be temporally derivable from multiple closed itemsets. Temporally closed itemsets form a (almost always proper) subset of closed itemsets in which temporal redundancies have been removed. The set of temporally closed itemsets can in fact be significantly smaller than the set of closed ones as we will demonstrate in our experimental evaluation in Section 8. At the same time, temporally closed itemsets are lossless in the sense that they can be used to uniquely determine the shape of the histories of all remaining itemsets.

7. Testing for Temporal Closedness

To check whether an itemset X is temporally non-closed we need to test whether an itemset XY exists which can be temporally derived from X. This, in turn, means we have to test whether ϵ in $\mathrm{supp}_i(XY) = \epsilon\,\mathrm{supp}_i(X), i = 1, \ldots, n$ is constant over time. Due to data usually being noisy as we showed in Figure 2, we will not check this criterion directly, but instead statistically test its validity. Also, we rewrite the criterion in an equivalent form to account for the order of values over time in the histories. Our experiments have shown that direct use of the criterion counterintuitively marked some histories as temporally derivable when they were noisy.

Let $\Delta_i \operatorname{supp}(X) := \frac{\operatorname{supp}_i(X)}{\operatorname{supp}_{i-1}(X)}$ be the relative change in support for itemset X between two periods T_{i-1} and $T_i, i = 2, \ldots, n$. Then, the above criterion holds, iff for any $i = 2, \ldots, n$

$$\Delta_i \operatorname{supp}(XY) = \Delta_i \operatorname{supp}(X) \quad (1)$$

This means, if the itemset XY is temporally derivable from X then the relative changes in the history of XY are equal to the temporally related relative changes in the history of the itemset X.

Imagine $\Delta_i \operatorname{supp}(X)$ and $\Delta_i \operatorname{supp}(XY)$ in a plotted graph, whereby $\Delta_i \operatorname{supp}(XY)$ is – as implied by Definition 2 – the dependent quantity. If $\Delta_i \operatorname{supp}(XY) = \Delta_i \operatorname{supp}(X)$ holds, then all points in the plot should be on a straight line with slope 1 and intercept 0. In practice, however, this equality will rarely hold due to noise. As a solution, we model the underlying relationship as $\Delta_i \operatorname{supp}(XY) = \Delta_i \operatorname{supp}(X) + \gamma$ where γ is a random error with zero mean and unknown, but low variance.

Under the assumption that the dependency of $\Delta_i \operatorname{supp}(XY)$ from $\Delta_i \operatorname{supp}(X)$ can be generally described by $\Delta_i \operatorname{supp}(XY) = a \cdot \Delta_i \operatorname{supp}(X) + b + \gamma$, we fit a regression line $\Delta \operatorname{supp}(XY) = \hat{a} \cdot \Delta \operatorname{supp}(X) + \hat{b}$. The parameters $\hat{a}$ and $\hat{b}$ are estimates for a and b and obtained by minimizing the regression error. We then test if $\Delta_i \operatorname{supp}(X)$ is statistically equal to $\Delta_i \operatorname{supp}(XY)$ by carrying out the following two steps:

1. Based on the estimates $\hat{a}$ and $\hat{b}$ we test the hypothesis that the true parameters of the model are $a = 1$ and $b = 0$ using a standard t-test [13].

2. Additionally, we test if the variance of γ is small, i.e. if $(\Delta_i \operatorname{supp}(X), \Delta_i \operatorname{supp}(XY))$ are sufficiently close to the regression line by setting a threshold $\tilde{r}$ for Pearson's correlation coefficient r.

Figure 3 illustrates the testing procedure. It shows the scatter plot of the relative changes of the support histories from Figure 1. The fitted regression line is $\Delta \operatorname{supp}(XY) = 1.0332 \cdot \Delta \operatorname{supp}(X) - 0.0396$ and the correlation coefficient $r \approx 0.9545$. The above test procedure using a significance level of 0.05 and $\tilde{r} = 0.95$ shows that XY is indeed temporally derivable from the history of X.

8. Experimental Results

As Theorem 1 as the central result of this publication states temporally closed itemsets form a subset of those itemsets which are closed over a sequence of time periods. For this reason, the question to be answered experimentally is how much the set of temporally closed itemsets is smaller than the set of closed itemsets.

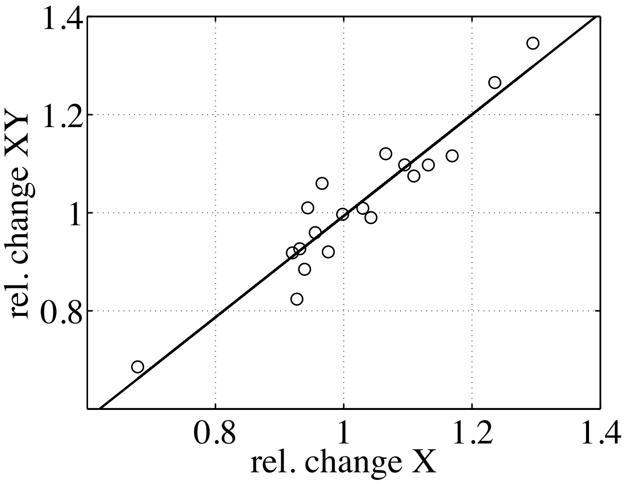

Figure 3. Scatter plot of the relative changes of the support histories shown in Figure 1. The fitted regression line is $\Delta \operatorname{supp}(XY) = 1.0332 \cdot \Delta \operatorname{supp}(X) - 0.0396$ and the correlation coefficient $r \approx 0.9545$.

For our experiments we chose two data sets. One data set, here called CRS, is extracted from the data-warehouse of a telecommunication company. The other data set we extracted from the IPUMS project[1] [16] which is dedicated to collecting, harmonizing and freely distributing census data.

The CRS data set contains answers of customers to a survey collected over a period of 20 weeks. Each record is described by 19 nominal attributes with a domain size between 2 and 9. We transformed the data set into a transaction set by recoding every (attribute, attribute value) combination as an item. Then we split the transaction set into 20 subsets, each corresponding to a period of one week. The subsets contain between 385 and 547 transactions. To each subset we applied a frequent itemset miner[2] using a minimum support threshold of $\operatorname{supp}_{\min} = 0.05$. From the obtained 20 sets of itemsets we only kept those itemsets which had been discovered in every period, i.e. those with complete support histories.

The data set we extracted from IPUMS contains census data of the USA collected during the years 2001–2006. Due to the data set being vast we restricted the data to the states New Jersey, New York, and Pennsylvania. From the available attributes we selected 15 concerning the person himself (e.g. age, race, gender), the house they are liv-

[1] http://usa.ipums.org/usa/

[2] We did use the frequent itemset miner contained within the *apriori* software package by Ch. Borgelt. It can be obtained from http://borgelt.net/fpm.html

ing in (e.g. number of bedrooms, year of built), and their profession (e.g. travel time, avg. hours worked per week, net income). Numeric attributes were converted into nominal ones using uniform binning. The domain size of the attributes varies between 2 and 9. We split the data set year-wise resulting in six data sets each containing between 130364 (for 2002) and 397788 (for 2006) records. We applied the same preprocessing and mining steps as for the CRS data.

We then tested the itemsets obtained from each data set for temporally closed itemsets by applying Definition 3 in combination with the test procedure in Section 7. We also tested for itemsets which are closed over the sequence of time periods. Here, we employed two approaches. The first one uses the original definition which requires strict equality of support values (cf. Definition 4). To rule out the effects of low quality data we also tested for *approximate closedness*, i.e. we did regard an itemset as non-closed if its support value is approximately the one of a more general itemset. Here, we did use the test from Section 7 extended by an additional test for $\epsilon > 0.98$ because for *strict closedness* it must be $\epsilon = 1$ (cf. Lemma 4).

The experimental results are shown in Table 1. Each row in the table corresponds to one data set, CRS or IPUMS. The column 'All' shows the number of all itemsets discovered, the following columns show how many of these itemsets are temporally closed itemsets, approximately closed itemsets and strictly closed itemsets, in this order. Both, absolute and relative numbers are given. As can be seen, the approach of temporally closed itemsets leads to a significant reduction in the number of itemsets compared to both conventional closed itemset approaches. While mining only for closed itemsets reduces the CRS result set to roughly 69% and the IPUMS result set to roughly 76% of its initial size, the temporally closed itemset approach leads to reduction of 36% and 24%, respectively. This means, for the CRS data the set of temporally closed itemsets is by a factor of 1.7 smaller than the set of strictly closed itemsets. For this IPUMS data this factor is with 3.1 even better. Figure 4 and Figure 5 show how the factor ϵ is distributed which maps the history of a non-temporally closed itemset to the smallest temporally closed itemset derivable from it. As we may expect from the results in Table 1 the range of ϵ is spread over a large range. The bar on the very right side in each histogram rougly indicates the number of itemsets that would have been discarded by a closed itemset approach. Because temporally closed itemsets are also closed itemsets the experimental results show that by exploiting temporal redundancies the set of closed itemsets can be further reduced by a very large extent, hence making it easier for a user to browse the discovered itemsets.

9. Conclusion and Future Research

Frequent itemset discovery suffers from the problem that typically a vast number of itemsets are generated. The large number makes them not only difficult to examine by a user but also influence the efficiency of subsequent processing steps. In the recent past, there have been considerable research efforts to exploit the time dimension in order to find novel ways to solve the relevance problem, no research had been done on how to utilize the temporal dimension in order to produce a reduced set of itemsets.

In this paper we introduced temporally closed itemsets as an extension to closed itemsets. In contrast to closed itemsets our approach also takes redundancies into account that are only visible if itemsets are observed over time. As the central theorem of this paper we proved that temporally closed itemsets are a subset of closed itemsets. Based on temporally closed itemsets it is possible to derive the shape of all other itemsets. Our experiments not only demonstrated that temporally closed itemsets do exists in real-world data. We also showed that the set of temporally closed itemsets can be smaller than the set of closed itemsets by a factor of two to three and by orders of magnitude smaller than the set of initially discovered itemsets.

As extensions to our work on temporally closed itemsets as presented in this paper, we are currently looking into the following two problems. First of all, we are working on a time and memory efficient algorithm to discover frequent temporally closed itemsets directly from data. Secondly, we aim to use temporally closed itemsets to generate a reduced set of association rules. Based on our current experiments it can be expected that they are significantly less in number than the results produced by other rule mining approaches.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993. ACM.

[2] R. Agrawal and G. Psaila. Active data mining. In M. Fayyad, Usama and R. Uthurusamy, editors, *Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3–8, Montreal, Quebec, Canada, 1995. AAAI Press.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *ICDE '95: Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[4] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations Newsletter*, 2(2):66–75, 2000.

[5] M. Boettcher, D. Nauck, D. Ruta, and M. Spott. Towards a framework for change detection in datasets. In M. Bramer,

Table 1. Experimental Results for the CRS and IPUMS data set

Data Set	All	Temp. Closed	Approx. Closed	Strictly Closed
CRS	1151	421 (36.5%)	715 (62.1%)	804 (69.8%)
IPUMS	3356	826 (24.6%)	2147 (63.9%)	2562 (76.3%)

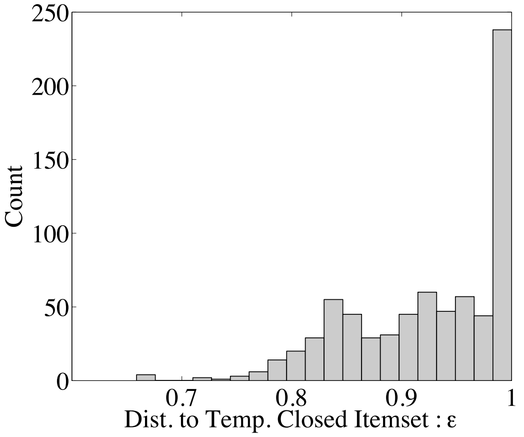

Figure 4. Histogram of the distance ε of non-temporally closed itemsets to the corresponding temporally closed one for the CRS data.

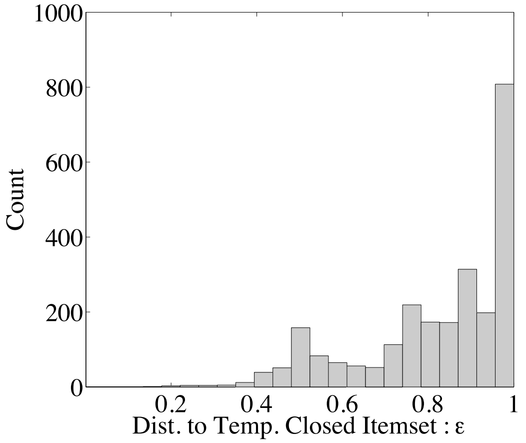

Figure 5. Histogram of the distance ε of non-temporally closed itemsets to the corresponding temporally closed one for the IPUMS data.

editor, *Research and Development in Intelligent Systems*, volume 23 of *Proceedings of AI-2006, the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 115–128. BCS SGAI, Springer, December 2006.

[6] M. Boettcher, M. Spott, and D. Nauck. Detecting temporally redundant association rules. In *Proceedings of the 4th International Conference on Machine Learning and Applications*, pages 397–403. IEEE Computer Society Press, 2005.

[7] A. Bykowski and C. Rigotti. A condensed representation to find frequent patterns. In *PODS '01: Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 267–273, New York, NY, USA, 2001. ACM.

[8] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 74–85, London, UK, 2002. Springer-Verlag.

[9] S. Chakrabarti, S. Sarawagi, and B. Dom. Mining surprising patterns using temporal description length. In *Proceedings of the 24th International Conference on Very Large Databases*, pages 606–617. Morgan Kaufmann Publishers Inc., 1998.

[10] S. Laxman, P. S. Sastry, and K. P. Unnikrishnan. A fast algorithm for finding frequent episodes in event streams. In *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 410–419, New York, NY, USA, 2007. ACM.

[11] B. Liu, W. Hsu, and Y. Ma. Discovering the set of fundamental rule changes. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 335–340, 2001.

[12] B. Liu, Y. Ma, and R. Lee. Analyzing the interestingness of association rules from the temporal dimension. In *Proceedings of the IEEE International Conference on Data Mining*, pages 377–384. IEEE Computer Society, 2001.

[13] D. Montgomery and G. Runger. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2002.

[14] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[15] J. Pei, J. Han, and L. V. S. Lakshmanan. Mining frequent item sets with convertible constraints. In *Proceedings of the 17th International Conference on Data Engineering*, pages 433–442, Washington, DC, USA, 2001. IEEE Computer Society.

[16] S. Ruggles, M. Sobek, T. Alexander, C. A. Fitch, R. Goeken, P. K. Hall, M. King, and C. Ronnander. Integrated public use microdata series: Version 4.0 [machine-readable database], Minneapolis, MN: Minnesota population center [producer and distributor], 2008.

[17] M. Spiliopoulou, S. Baron, and O. Gnther. Efficient monitoring of patterns in data mining environments. In *Pro-*

ceedings of the 7th East-European Conference on Advances in Databases and Information Systems (ADBIS'03), pages 253–265. Springer, September 2003.

[18] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[19] M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining*, pages 457–473. SIAM, 2002.