A Condensed Representation of Itemsets for Analyzing Their Evolution over Time

Mirko Boettcher[1], Martin Spott[2], and Rudolf Kruse[1]

[1] University of Magdeburg, Faculty of Computer Science,
39106 Magdeburg, Germany
{miboettc,kruse}@iws.cs.uni-magdeburg.de
[2] Intelligent Systems Research Centre, BT Group plc,
Adastral Park, Ipswich IP5 3RE, United Kingdom
martin.spott@bt.com

Abstract. Driven by the need to understand change within domains there is emerging research on methods which aim at analyzing how patterns and in particular itemsets evolve over time. In practice, however, these methods suffer from the problem that many of the observed changes in itemsets are temporally redundant in the sense that they are the side-effect of changes in other itemsets, hence making the identification of the fundamental changes difficult. As a solution we propose temporally closed itemsets, a novel approach for a condensed representation of itemsets which is based on removing temporal redundancies. We investigate how our approach relates to the well-known concept of closed itemsets if the latter would be directly generalized to account for the temporal dimension. Our experiments support the theoretical results by showing that the set of temporally closed itemsets is significantly smaller than the set of closed itemsets.

1 Introduction

In many application areas data is being collected over a long time. Due to its temporal nature such data not only captures influences, like market forces or the launch of a new product, but also reflects the changes of the underlying domain. Often, change can mean a risk (like a shrinking subgroup of target customers) or an opportunity (like an evolving market niche). In either case, it is in many domains not only imperative to detect change in order to survive or to win but it is also essential for successful decision making to analyze and act upon it.

As a response to this need there is increasing research interest in methods which aim at analyzing the changes within a domain by describing and modeling how the results of data mining—models and patterns—evolve over time. The term change mining has been coined as an umbrella term for such methods [1]. Change mining approaches have been proposed for a variety of patterns and models. Nonetheless, many studies focus on analyzing change in the context of itemsets, not only because itemsets are rather comprehensible itself but also because their evolution can be represented in a convenient and interpretable way.

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 163–178, 2009.

It is generally assumed that itemsets cannot change in their symblic representation but only in quantitative measures which describe them, the most common of which is support, i.e. the frequency of occurrence within a data set. The evolution of itemsets is therefore captured in time series (also called histories), whereby most approaches only utilize support histories [2,3,4,5,6]. Nevertheless, they can often be adapted to other measures.

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 change mining is applied to this dataset to discover evolving customer segments in a sociographical context. Assume that the following itemset which specifies one customer segment has been discovered: XY : Broadband=Yes, Gender=Male. Figure 1 shows its support history describing the size of this segment relative to all customers over 20 periods. Change mining approaches would, for example, employ statistical tests [4,6], pattern matching in time series [2] or heuristics [5] to detect that this history shows many characteristic features which might be of interest to the expert: a trend turning point and a declining and inclining trend to the left, respectively to the right of it.

Generally, changing itemsets hint at changes in the underlying domain. Such changes may indicate that an intervening action is required, for instance, to rectify a problem [3]. On the other hand, an itemset which always remains stable in its support can be expected to describe an *invariant* of the domain. Invariants, however, are of less interest to experts because they are almost always known and usually do not indicate a serious problem [7]. Still, they may be of interest if the domain under consideration is in its basic underlying principles has not been fully understood, yet.

Besides being only of secondary interest, invariants also lead to a drastically larger number of itemsets with changes in their histories. To continue our example, assume that the fraction of males among all broadband users is an invariant of the domain. In this case, the change of X : Broadband=Yes will be qualitatively the same as the one of XY : Broadband=Yes, Gender=Male. Both histories will only differ by a scaling factor which, in turn, is the invariant. This is also shown in Figure 1. In the context of change mining it can be said that both itemsets are *temporally redundant* with respect to each other.

Temporal redundancy is very common in practice. Many of the changes observed in itemset histories are simply the effect of a combination of changes in other itemset histories with domain invariants. For an expert it will be a very tedious task to identify the fundamental changes in which he is primarily interested. This problem is even worsened by the vast number of itemsets which are discovered. For this reason it is desirable to obtain a *condensed representation* which captures the fundamental set of changing itemset histories and allows to reconstruct the shape of all other itemsets histories that are necessary for change mining. To our knowledge this problem has up to now neither been addressed in the field of change mining nor in the area of condensed representations of itemsets.

Our goal in this paper is twofold: first of all we want to contribute to the field of change mining by providing a condensed representation of itemsets which incorporates the temporal dimension. More precisely, we introduce *temporally closed itemsets* as an approach to reduce the number of itemsets by accounting for temporal redundancies. The set of temporally closed itemsets is minimal in the sense that the shape of every other itemset's history can be reconstructed from it. Here we follow the argument of Agrawal and Psaila [2] and Chakrabarti et al [3] that an itemset's relevance is primarily dictated by its qualitative change over time and not by its actual support value. Secondly, we want to tighten the link between itemset mining and itemset change mining by investigating how temporally closed itemsets relate to the well-known concept of closed itemsets. In particular, we show that the set of temporally closed itemsets is a subset of the set of closed itemsets if the notion of the latter is directly generalized to the temporal dimension. The subset property is the result of removing those redundancies from the set of closed itemsets which are only visible when itemsets are analyzed over time.

The remainder of this paper is organized as follows. In Section 2 we discuss related work. Section 3 and Section 4.1 introduce the necessary background on frequent itemset mining and closed itemsets. In Section 4.2 we discuss how closed itemsets can be straightforwardly generalized to be applicable to sequences of time periods in order to have a benchmark for our approach. 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 how the set of temporally closed itemsets can be discovered. 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: change mining and condensed representations. For this reason we will first provide an overview over existing change mining methods for associations, followed by some background on condensed representations.

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 [4] showed how trend, semi-stable and stable rules can be distinguished using a statistical approach. In [3] 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 [5,1]. 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. Liu et al proposed a method to detect so-called *fundamental rule changes* that aims to

identify changes in support and confidence of association rules which cannot be explained by other changes [8]. The authors provide heuristic criteria for solving this task. However, their approach differs to our approach of temporally closed itemsets because it 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.

Several approaches have been proposed which lead to a condensed representation of the set of discovered itemsets such that all other itemsets can be derived from the representation. Three such techniques are: closed itemsets [9,10], counting inference [11] and deduction rules [12]. 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. We will discuss them in more detail in Section 4.1.

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 $\mathrm{supp}_{\min}$. 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.

As other authors we define the change of an itemset 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 $\mathrm{supp}_i(X)$. 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 $(\mathrm{supp}_1(X), \ldots \mathrm{supp}_n(X))$ which are also called *support histories*.

4 Closed Itemsets

4.1 Definition

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 [9]. 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. This means that non-closed sets can be regarded redundant.

Formally, a closed itemset is defined as follows (cf. [9]):

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

Equivalently, an itemset X is called *non-closed* iff there exists a proper superset Y such that $\text{supp}(X) = \text{supp}(Y)$. The largest of such supersets is also called the *closure* of X.

As already mentioned in Section 2 closed itemsets can only be applied to a single data set. As a result, they do not account for redundancies imposed by the temporal dimension, where we deal with a sequence of data sets.

4.2 Generalization to a Sequence of Time Periods

To analyze the approach proposed in this paper it is desirable to have an established condensed representation with which it can be compared, both theoretically and experimentally. Although closed itemsets are probably the most widely used condensed representation they cannot directly be used for this purpose because they were not developed with the temporal dimension in mind. For this reason the definition of closed itemsets given in the previous section has to be generalized from a single data set corresponding to one time period to a sequence of time periods.

On the one hand, from such a generalization one would expect that it is straightforward in the sense that it resembles the original definition of closed itemsets as good as possible. On the other hand, the generalization should be sufficient for the purpose of change analysis: an itemset should only be regarded as *non-closed over a sequence of time periods* if no change information is lost.

As an example how change information can be lost consider a non-closed itemset whose closure varies across periods. Should this non-closed itemset be regarded as closed or as non-closed over a sequence of time periods? As laid out in the Introduction a notion of redundancy in the context of change analysis should be based on invariants of the domain. If an itemset with varying closure would be regarded as non-closed over a sequence of time periods and thus not presented to a user this change information is lost. Consequently, it is reasonable to regard itemsets whose closure varies across periods as closed in the context of a generalization of closed itemsets towards sequences of time periods.

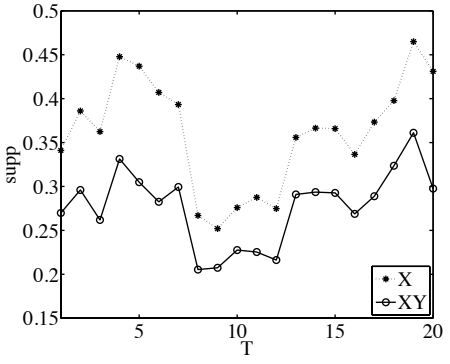

Fig. 1. Histories of the itemsets XY and X showing that $X \hookrightarrow XY$

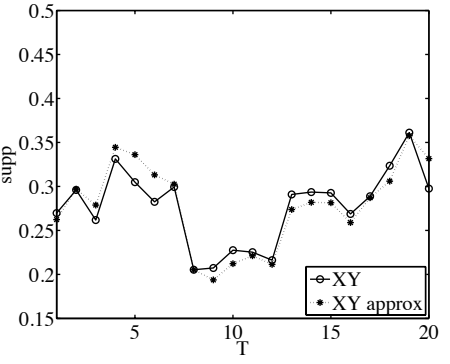

Fig. 2. Approximated history of XY using the history of X

Generally, two requirements have to be met by a temporal generalization of closed itemsets. In the first place, an itemset X should be non-closed over a sequence of time periods only if it is non-closed in all periods. Secondly, following the discussion in Section 1 the underlying reason for non-closedness should be an invariant, i.e. the superset Y with equal support (the closure Y of X) should be the same in every time period.

A direct temporal generalization of non-closed itemsets which meets the requirements above is the following: An itemset is *non-closed over a sequence of time periods* $T_i, i = 1, \ldots, n$ iff there exists an itemset $Y \supset X$ such that for all periods $\text{supp}_i(X) = \text{supp}_i(Y)$ $i = 1, \ldots, n$. Equivalently, itemsets which are *closed over a sequence of time periods* are defined as follows:

Definition 2 (Closed over a Sequence). *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* $\text{supp}_i(X) = \text{supp}_i(Y), i = 1, \ldots, n$.

The above definition means that the set of itemsets which are closed over a sequence of time periods contains all itemsets which are closed in at least one period. Additionally, it also contains itemsets X which are non-closed in each individual time period but for which their closure differs across periods. It should be noted, however, that the latter is an extremely rare constellation as it has already been reported by authors in the field of incremental itemset mining (cf. [13]). Indeed, we did not observe non-closed itemsets with temporally varying closure in the data sets we used for our experiments.

If the context is clear we will for brevity refer to the *direct temporal generalization of closed itemsets* introduced in this section simply as closed itemsets.

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

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 $X \overset{\epsilon}{\hookrightarrow} Y$. From the criterion $\text{supp}_i(XY) = \epsilon\, \text{supp}_i(X), i = 1, \ldots, n$ used within the definition it directly follows that ϵ decreases with increasing size of Y, i.e. for $X \overset{\epsilon_1}{\hookrightarrow} Y$ and $X \overset{\epsilon_2}{\hookrightarrow} Z$ with $Y \subset Z$ it is $\epsilon_2 \leq \epsilon_1$. The criterion $\text{supp}_i(XY) = \epsilon\, \text{supp}_i(X), i = 1, \ldots, n$ can also be rewritten as $\epsilon = \text{supp}_i(XY)/\,\text{supp}_i(X) = P(XY \mid T_i)/P(X \mid T_i) = P(Y \mid XT_i)$. 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 7.2 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. Lemma 1 follows directly from Definition 3.

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

6 Temporally Closed Itemsets

If XY is temporally derivable from X then XY and Y have histories with qualitatively the same shape. Further, if we assume that the relevance of an itemset is primarily determined by the qualitative changes represented in its history both itemsets, X and XY, would have the same interestingness. 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 can be regarded as temporally 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 4 (Temporally Closed Itemset). *An itemset X is temporally closed iff there exists no itemset $Y \supset X$ such that $X \hookrightarrow Y$.*

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.

To analyze how temporally closed itemsets relate to the temporal generalization of closed itemsets discussed in Section 4.2 the definition of an itemset which is closed over a sequence of time periods needs to be linked to the notion of temporal derivability. By comparing Definition 2 with Definition 3 it can be seen that the link between the two concepts can be expressed as follows:

Lemma 3. *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. Follows directly from the definition of a temporally derivable itemset (cf. Definition 3).

We now have the necessary tools to prove the central theorem of this paper which shows that temporally closed itemsets are a subset of the direct temporal generalization of closed itemsets introduced in Section 4.2, i.e. a subset of the itemsets which are closed over a sequence of time periods.

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.

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

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

The following counterexample shows that TC can indeed be 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 2 it is $X_1 \hookrightarrow X_4$ and $X_3 \hookrightarrow X_4$. Using Definition 4 it follows that $X_1 \notin TC$ and $X_3 \notin TC$.

This means, every temporally closed itemset is also closed over a sequence of time periods but not every itemset which is closed over a sequence is also a temporally closed one. The counterexample shows that a temporally closed itemset can be temporally derivable from multiple closed itemsets. As we will see in our experiment results in Section 8 temporally closed itemsets form a (almost always proper) subset of directly generalized 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 Discovery Procedure

To obtain the set of temporally closed itemsets we use a two step approach in which we first generate a set of candidate itemsets and then test every candidate whether it is temporally closed, or not. This two step procedure will be detailed in the following.

7.1 Candidate Generation

A naive approach to obtain a candidate set would be to consider the set of itemsets which are frequent in every time period. Because this set is usually vast the subsequent testing step would be very time consuming. A more efficient approach is to restrict the candidates to the set C of frequent itemsets which are closed over a sequence of time periods. According to Theorem 1 this set is a superset of the set of temporally closed itemsets TC. It is, however, by several factors smaller than the set of all frequent itemsets.

Given a time-stamped data set D and a segmentation $T_i := [t_{i-1}, t_i]$ of the covered time span $[t_0,\ t_n]$, D is divided into the corresponding subsets $D_i \subset D$. From this sequence of temporally ordered, disjunct data sets $D_1, \ldots, D_n$ the candidate set C can be obtained in three steps.

Mining Closed Itemsets in each Period. An algorithm for closed itemset mining is applied to each data set which yields a sequence of closed itemset sets $C_1, \ldots, C_n$.

Merging the Sets of Closed Itemsets. The closed itemset sets $C_1, \ldots, C_n$ have to be inserted into the candidate set C. We start by setting $C := C_1$ and converting C into a data structure in which each node represents an itemset and maintains a list of parents (largest subsets) and children (smallest supersets). We then subsequently insert the remaining C_i in C whereby the list of parents and childrens in each node is being maintained. Assume that $C_1, \ldots, C_{j-1}$ have already been inserted in C and that $X \in C_j$ is the next itemset to be inserted. The following scenarios are possible:

- *X already exists in C:* In this case the support of X in the data set D_j is inserted at position j in the support history of X in C.
- *X does not exist in C:* In this case it is checked whether there exists a $Y \in C$ such that $X \subset Y$. If no such proper superset exists, X was infrequent in earlier periods and is not inserted because a complete support history of it cannot be obtained anymore. If a proper superset exists X was frequent but non-closed in earlier periods and thus is inserted into C. The support of X in the data set D_j is inserted at position j in the support history of X in C.

Complexity. Assume that m is the length of the longest itemset, that the number of different items is p, and that all C_i have equal size. The initial computational effort to create the data structure for C is $O(|C|^2)$. The effort to search an itemset X in C is bound by the length m of the longest possible path in C. To insert an itemset X into C after the smallest proper superset has been found all largest proper subsets and all smallest proper supersets of X in C need to be accessed and their parent list, respectively children list, updated. This can be accomplished in $O(m+p)$ because every itemset cannot have more than m parent nodes and more than p child nodes. Under the assumption that always a fraction α of itemsets in C_i needs to be inserted into C the overall complexity of inserting the sets $C_2, \ldots, C_n$ into C therefore is $O(|C_i|^2 + mn|C_i| + \alpha n(m+p)|C_i|)$. A bottleneck is the quadratic effort for creating the initial data structure. This effort, however, can be avoided by using in the previous step a closed itemset miner like CHARM-L [10] which outputs the data structure directly.

Handling Incomplete Histories. It needs to be checked whether C contains itemsets with incomplete histories. An incomplete history can occur either because an itemset was non-closed or infrequent in some but not all periods, respectively. We iterate over each element $X \in C$ starting with the largest itemsets. We visit itemsets with no children always first and visit all others in order of size. The following two scenarios for itemsets X with incomplete histories are possible:

- *The itemset X has no children:* This implies that in at least one period either X or its closure were infrequent. In this case it is not possible to obtain a

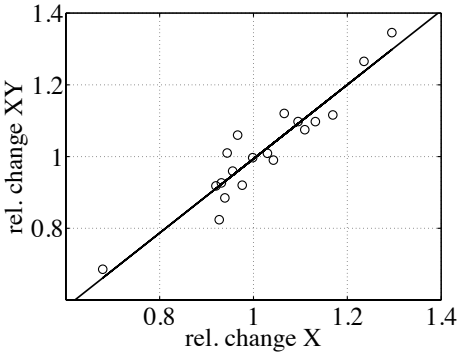

Fig. 3. Scatter plot of the relative changes of the support histories shown in Figure 1. The fitted regression line is $\Delta\,\mathrm{supp}(XY) = 1.0332 \cdot \Delta\,\mathrm{supp}(X) - 0.0396$ and the correlation coefficient $r \approx 0.9545$.

complete support history of X, thus the itemset is removed from C and the children list of its parents updated accordingly.

- *The itemset X has children:* This implies that the itemset X was non-closed in at least one period. Assuming that its support is missing in period j then one of the children of X must have a support value for this period, either because it is the closure of X in j or because the support value was completed in an earlier step. Due to X being non-closed in at least one period, there may exist subsets $Z \subset X$ which are non-closed in every period but have a different closure across periods (one of which is X). Following the discussion in Section 4.2 these itemsets are closed over a sequence of time periods and thus also part of the candidate set C. For this reason, each itemset $Z \subset X$ for which no itemset $X' \in C$ exists such that $X' \subset Z \subset X$ is also inserted into C.

Complexity. We use the same notation as in the complexity analysis before. Additionally we assume that a fraction β of itemsets in C is infrequent and that a fraction γ is non-closed in some but not all periods, respectively. In the first scenario for each deleted itemset the children list of each parent needs to be updated. Because each itemset can have at most m parents the computational effort thus is $O(m\beta|C|)$. In the second scenario, for each of the $\gamma|C|$ itemsets each direct child needs to be accessed. Since the number of children is bound by p the computational effort therefore is $O(p\gamma|C|)$. Further the set NC of itemsets which were non-closed in each period but had a varying closure need to be inserted. As in the previous step the complexity of inserting an itemset is $O(m+p)$. Taking into account that every itemset with complete history needs to be accessed once the overall complexity therefore is $O((1+(m-1)\beta+(p-1)\gamma)|C|+(2p+m)|NC|)$. We do not have an estimation for β, γ and $|NC|$ but in our experiments we neither encountered itemsets with varying closedness nor with varying closure. This gives rise to the assumption that γ and $|NC|$ will be very small in practice. Concerning β we observed that typically $10-20\%$ of the itemsets in C had an incomplete history due to infrequency in some periods.

7.2 Testing for Temporal Closedness

To check whether an itemset $X \in C$ 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 $\Delta_i \,\mathrm{supp}(X) := \frac{\mathrm{supp}_i(X)}{\mathrm{supp}_{i-1}(X)}$ 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 $\Delta_i \,\mathrm{supp}(XY) = \Delta_i \,\mathrm{supp}(X)$ for any $i = 2, \ldots, n$. 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 $\Delta_i \,\mathrm{supp}(X)$ and $\Delta_i \,\mathrm{supp}(XY)$ in a plotted graph, whereby – as implied by Definition 3 – $\Delta_i \,\mathrm{supp}(XY)$ is the dependent quantity. If $\Delta_i \,\mathrm{supp}(XY) = \Delta_i \,\mathrm{supp}(X)$ 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 $\Delta_i \,\mathrm{supp}(XY) = \Delta_i \,\mathrm{supp}(X) + \gamma$ where γ is a random error with zero mean and unknown, but low variance.

Under the assumption that the dependency of $\Delta_i \,\mathrm{supp}(XY)$ from $\Delta_i \,\mathrm{supp}(X)$ can be generally described by $\Delta_i \,\mathrm{supp}(XY) = a \cdot \Delta_i \,\mathrm{supp}(X) + b + \gamma$, we fit a regression line $\Delta \,\mathrm{supp}(XY) = \hat{a} \cdot \Delta \,\mathrm{supp}(X) + \hat{b}$. The parameters $\hat{a}$ and $\hat{b}$ are estimates for a and b and obtained by minimizing the regression error. We then test if $\Delta_i \,\mathrm{supp}(X)$ is statistically equal to $\Delta_i \,\mathrm{supp}(XY)$ 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.
2. Additionally, we test whether the variance of γ is small, i.e. whether the $(\Delta_i \,\mathrm{supp}(X), \Delta_i \,\mathrm{supp}(XY))$ 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 $\Delta \,\mathrm{supp}(XY) = 1.0332 \cdot \Delta \,\mathrm{supp}(X) - 0.0396$ 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.

Complexity. The complexity of this step is apparently $O(|C|)$ because each itemset's parent can be directly accessed through the parent list.

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. The set of itemsets which are closed over a sequence of time periods also forms the candidate set to be tested for temporal closedness. For this reason, the question to be answered experimentally is how much the set of temporally closed itemsets is smaller than the set of itemsets which are closed over a sequence.

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] [14] 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.

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 living 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 steps as for the CRS data.

For each data set we then obtained the candidate set to be tested for temporal closedness. To each subset of each data set we applied a closed itemset miner[2] using 11 different minimum support thresholds in steps of 0.01 in the range from $\text{supp}_{\min} = 0.05$ to $\text{supp}_{\min} = 0.15$. For each value of $\text{supp}_{\min}$ we then generated from the obtained 20 sets of closed itemsets for CRS, respectively 6 sets for IPUMS, the candidate sets as described in Section 7.

For both data sets we observed that the closed itemsets were the same in every period, i.e. closed itemsets did not turn into non-closed ones and vice versa. This indicates that such events are very rare. It also implies that for both data sets the candidate set is equal to the set of closed itemsets in each period.

We tested the elements of the candidate sets for temporally closed itemsets by applying Definition 4 in combination with the test procedure in Section 7.2. We also investigated the number of itemsets which are closed over the sequence of

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

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

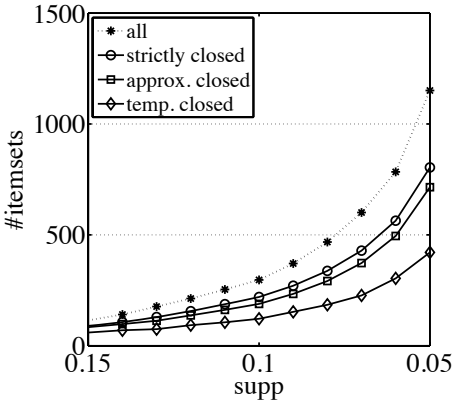

Fig. 4. Results for the CRS data set

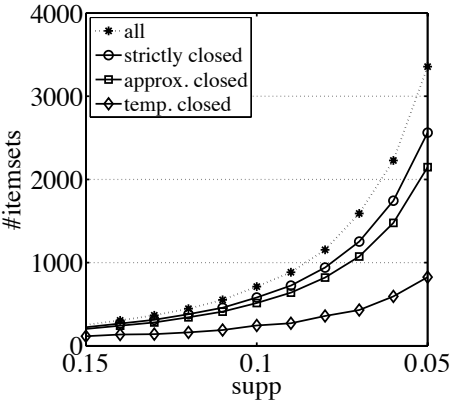

Fig. 5. Results for the IPUMS data set

time periods. Here, we employed two approaches. The first one uses the original definition which requires strict equality of support values (cf. Definition 2). To rule out the effects of low quality data we also tested for *approximate closedness*, i.e. we regarded an itemset as non-closed if its support value is approximately the one of a more general itemset. Here, we applied the test from Section 7.2 to the candidate set extended by an additional test for $\epsilon > 0.98$ because for *strict closedness* it must be $\epsilon = 1$ (cf. Lemma 3).

To compare the number of itemsets returned by our approach with the overall number of frequent itemsets (i.e. closed and non-closed frequent itemsets) we also applied a frequent itemset miner to both data sets using the same support threshold as for the other experiments. We only kept those itemsets which were frequent in every period.

The experimental results for the CRS data set are shown in Figure 4 and for the IPUMS data set in Figure 5 . Both figures show the the number of all frequent itemsets discovered and the number of itemsets which are temporally closed, approximately closed, and strictly closed. As can be seen, the approach of temporally closed itemsets leads to a significant reduction in the number of itemsets compared to both closed itemset approaches. For example, for $\text{supp}_{\min} = 0.05$ 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 while the temporally closed itemset approach leads to a 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 the IPUMS data this factor is with 3.1 even better.

Figure 6 and Figure 7 show for $\text{supp}_{\min} = 0.05$ 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 Figure 4 and Figure 5 the range of ϵ is spread over a large range approximately $[0.75, 1]$ for the CRS data and $[0.4, 1]$ for the IPUMS data. The bar on the very right side in

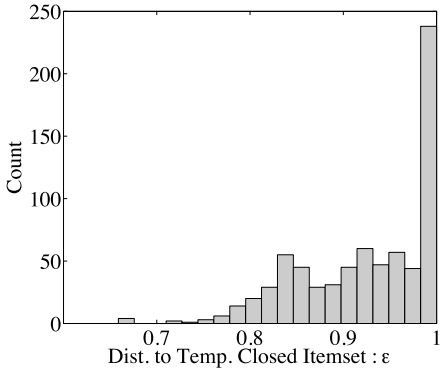

Fig. 6. Histogram of the distance ε of non-temporally closed itemsets to the corresponding temporally closed one for the CRS data.

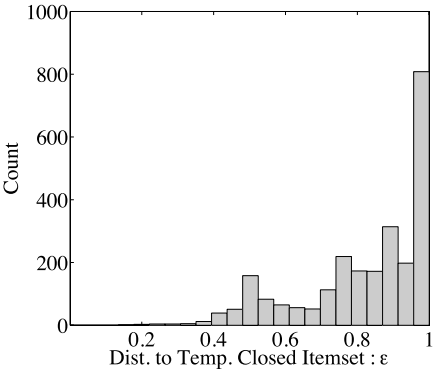

Fig. 7. Histogram of the distance ε of non-temporally closed itemsets to the corresponding temporally closed one for the IPUMS data.

each histogram ($\epsilon \approx 1$) roughly indicates the number of itemsets that would have been discarded by the generalized closed itemset approach described in Section 4.2. Our approach, in contrast, would discard every itemset shown in the histogram and thus reduce the set of closed itemsets by a very large extent.

9 Conclusion

Many businesses collect huge volumes of time-stamped data about all kinds of processes. This data reflects changes in the underlying domain. It is crucial for the success of most businesses to detect these changes, and finally to adapt or react to them. As a response to this need there is emerging research on data mining methods which aim at understanding change within a domain by analyzing how patterns evolve over time. Several studies have been conducted on analyzing how itemsets change over time. However, these approaches do not account for temporal redundancies, i.e. the problem that many of the observed changes are simply the side-effect of other changes.

In this paper we solved this problem by introducing temporally closed itemsets as a condensed representation of itemsets which is, on the one hand, free of temporal redundancies but which, on the other hand, still contains all the information needed for change analysis. Based on temporally closed itemsets it is possible to derive the shape of the history of all other itemsets. In particular, we showed that temporally closed itemsets are a subset of the set of closed itemsets if the definition of the latter would be directly generalized to be applicable to sequences of time periods. Our experiments not only demonstrated that temporally closed itemsets do exist 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.

References

1. Böttcher, M., Spiliopoulou, M., Höppner, F.: On exploiting the power of time in data mining. SIGKDD Explorations Newsletter 10(2), 3–11 (2008)
2. Agrawal, R., Psaila, G.: Active data mining. In: Fayyad, U.M., Uthurusamy, R. (eds.) Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Montreal, Quebec, Canada, pp. 3–8. AAAI Press, Menlo Park (1995)
3. Chakrabarti, S., Sarawagi, S., Dom, B.: Mining surprising patterns using temporal description length. In: Proceedings of the 24th International Conference on Very Large Databases, pp. 606–617. Morgan Kaufmann Publishers Inc., San Francisco (1998)
4. Liu, B., Ma, Y., Lee, R.: Analyzing the interestingness of association rules from the temporal dimension. In: Proceedings of the IEEE International Conference on Data Mining, pp. 377–384. IEEE Computer Society Press, Los Alamitos (2001)
5. Spiliopoulou, M., Baron, S., Günther, O.: Efficient monitoring of patterns in data mining environments. In: Kalinichenko, L.A., Manthey, R., Thalheim, B., Wloka, U. (eds.) ADBIS 2003. LNCS, vol. 2798, pp. 253–265. Springer, Heidelberg (2003)
6. Böttcher, M., Spott, M., Nauck, D., Kruse, R.: Mining changing customer segments in dynamic markets. Expert Systems with Applications 36(1), 155–164 (2009)
7. Berger, C.R.: Slippery slopes to apprehension: Rationality and graphical depictions of increasingly threatening trends. Communication Research 32(1), 3–28 (2005)
8. Liu, B., Hsu, W., Ma, Y.: Discovering the set of fundamental rule changes. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 335–340 (2001)
9. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Information Systems 24(1), 25–46 (1999)
10. Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Transactions on Knowledge and Data Engineering 17(4), 462–478 (2005)
11. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhai, L.: Mining frequent patterns with counting inference. SIGKDD Explorations Newsletter 2(2), 66–75 (2000)
12. Calders, T., Goethals, B.: Mining all non-derivable frequent itemsets. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 74–85. Springer, Heidelberg (2002)
13. Chi, Y., Wang, H., Yu, P.S., Muntz, R.R.: Catch the moment: maintaining closed frequent itemsets over a data stream sliding window. Knowledge and Information Systems 10(3), 265–294 (2006)
14. Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M., Ronnander, C.: Integrated public use microdata series: Version 4.0, machine-readable database. Minnesota population center, Minneapolis (producer and distributor) (2008)