Data Science and Pattern Recognition ©2021 ISSN 2520-4165
Ubiquitous International Volume **5**, Number **1**, March **2021**

Towards Predicting the Occurrence Times of Emerging and Decaying Patterns

Vincent Mwintieru Nofong

Faculty of Engineering
University of Mines and Technology
vnofong@umat.edu.gh

Hamidu Abdel-Fatao

Faculty of Engineering
University of Mines and Technology
habdel-fatao@umat.edu.gh

Abstract. *Emerging and decaying patterns are patterns whose occurrence frequencies vary (increase or decrease) from one database to another. They are essential in classification and trend prediction because they can reveal useful trends and contrasts in databases for decision making. Most studies to date, only focus on their discovery for classification with a few focusing on their application in trend prediction. Though recently some works showed their application in predicting trends in temporal (time-stamped) databases, such works are only able to predict their continuous emergence and decay but not their likely occurrence times. To enable the prediction of the likely occurrence times of emerging and decaying patterns in temporal databases, this work incorporates the concept of periodicity on an existing emerging and decaying pattern-based trend prediction model in predicting their likely occurrence times. Experimental analysis on real-world datasets show that with the incorporated periodicity, the likely occurrence times of emerging and decaying patterns in temporal databases can be accurately predicted for decision making.*
Keywords: Periodicity, Trend Prediction, Frequent Patterns, Emerging Patterns, Decaying Patterns.

1. **Introduction.** Frequent itemset (pattern) mining [2] [13] [36] [37] is a fundamental data mining task (with a wide range of applications) that has been widely studied over the past years. The goal in frequent pattern mining is to identify all patterns that occur frequently in a given database. For any given database, a pattern is said to be frequent if its frequency of occurrence (often referred to as support) within the database is not less than a user specified threshold. Mining frequent patterns is not a challenging task since the support of patterns follow the anti-monotone property. That is, for any pattern that is frequent, all its subsets will also be frequent since its support will be less than or equal to the support of its subsets. Similarly, for a pattern that is infrequent, all its supersets will also be infrequent since its support will be greater than or equal to the supports of its supersets. This anti-monotonic nature of patterns support thus help in reducing the search space in frequent pattern mining.

Research on discovery of emerging patterns was introduced by Dong and Li [3], wherein emerging patterns are referred to as itemsets whose supports increase significantly from one database (dataset), $\mathbf{D_1}$, to another $\mathbf{D_2}$. Specifically, they define an emerging pattern as an itemset whose growth rate (which is, the ratio of its support in $\mathbf{D_2}$ over $\mathbf{D_1}$), is no less than a user-defined threshold.

Emerging pattern mining has since been a major data mining task in various decision support systems as they often reveal hidden and useful contrasts or trends in databases. Various works on discovering emerging patterns over the past years have shown their effective use in: classification [5] [8] [17] [20], understanding gene regulations [18] [19], understanding behaviour [21] [31] [35], extracting patterns in data streams [12] and spatiotemporal data [22], and trend prediction [23] [24] [25] [26]. Discovery of emerging patterns is however a challenging task as the emergence of patterns is neither monotonic nor anti-monotonic [3] [30]. As such, it makes it difficult to terminate the emerging pattern mining process as a pattern's emergence, unlike its frequency, cannot be used in reducing the search space in the early stage of emerging pattern mining.

Several techniques and constraints have been proposed in recent years to mine categories of emerging patterns for various decision making. Typical of such techniques and constraints can be found in works on mining jumping emerging patterns [6] [8] [16] [34] and interesting emerging patterns [7]. Notwithstanding the several techniques and constraints for mining various types of emerging patterns, as pointed out in [24] [25] [26], emerging pattern discovery still faces a number of challenges such as: reporting too many emerging patterns, missing some useful emerging patterns, and reporting redundant or trivial emerging patterns.

Besides these challenges, most works focus on the discovery of emerging patterns in static databases for classification with a few [23] [24] [25] [26] recently focusing on their application in trend prediction. Though works such as [23] [24] [25] [26] recently researched on and proposed techniques for trend prediction with emerging and decaying patterns, their techniques which are able to predict the continuous emergence or decay of patterns in temporal (time-stamped) databases, are unable to predict their likely occurrence times within the databases.

The work in this paper is motivated by the inability of existing emerging and decaying pattern trend prediction techniques in predicting the likely occurrence times of patterns. To enrich trend prediction using emerging and decaying patterns, and to be able to predict their likely occurrence times in temporal databases, the concept of periodicity from periodic frequent pattern mining [1] [9] [10] [11] [15] [27] [28] [29] [33] is incorporated on an existing emerging and decaying pattern-based trend prediction model to predict their likely occurrence times.

This paper makes the following contributions in emerging and decaying pattern mining:

1. It shows an application of emerging and decaying patterns for trend prediction.
2. It shows that by incorporating the concept of periodicity (from periodic frequent pattern mining) in trend prediction using emerging and decaying patterns, their likely occurrence times in time-stamped databases can be effectively predicted.

The rest of the paper is as follows. Section 2 presents the related work on emerging pattern and periodic frequent pattern mining. Section 3 presents the definitions and problem statement while Section 4 presents the approach of incorporating periodicity in trend prediction with emerging and decaying patterns. Section 5 presents the experimental analysis while Section 6 presents the conclusions and future works.

2. **Related Work.** Emerging pattern mining and their applications, since the introduction of the concept by Dong and Li [3], has generated significant research interest especially in decision support systems. Evidently, several studies have emerged in this research domain among which are works such as [4] [6] [7] [8] [16] [18] [32] [34].

Some researchers [6] [7] [8] [16] however argue that the emerging pattern definition proposed in [3] often generates too many emerging patterns hence making it difficult identifying the set of interesting and useful patterns to support decision making. Several techniques and constraints have since been proposed to enable mining of categories of emerging patterns to support domain specific decision making. Typical of such works include: essential emerging patterns [6], jumping emerging patterns [8] [16] [34] and interesting emerging patterns [7].

Though the above mentioned works have been useful in mining emerging patterns for various decision making, they often report a too large or a too small number of emerging patterns. Additionally, they majorly focus on discovering interesting emerging patterns from static databases for classification and seldom discuss their discovery in time-stamped databases to support trend prediction.

Recent works [23] [25] [26] which researched on the applications of emerging patterns for trend prediction observed that, in time-stamped databases, most existing emerging patterns will not be suitable for trend prediction as they could:

i) be noisy, spiky or false trends (typical of emerging patterns reported in [8] [16] [34])
ii) contain redundant emerging patterns (typical of emerging patterns reported in [3])
iii) miss some important emerging patterns (typical of emerging patterns reported in [7] [8] [16] [34]), or,
iv) be emerging or decaying by random chance without inherent item relations (typical of emerging patterns reported in [8] [16] [34]).

To mine emerging patterns in time-stamped databases that will be suitable for trend prediction, Nofong *et al.* [26] proposed the non-derivable emerging patterns (emerging patterns that do not contain redundant information). Nofong [25] also proposed the productive emerging patterns (emerging patterns that have inherent item-relations and not due to random chance or data fluctuations) also for trend prediction. Based on their proposed types of emerging patterns, the works in [25] and [26] showed using naive techniques the effective use of emerging patterns for trend prediction in time-stamped databases. Though the techniques in [25] and [26] are able to predict the continuous emergence or decay of patterns in time-stamped databases, they are unable to predict their likely supports for decision making. To enable the prediction of the supports emerging and decaying patterns in time-stamped databases, [24] proposed EDTrend, a methodology for trend prediction with emerging and decaying patterns.

Though the proposition in [24] has been shown to be effective in predicting the likely supports of emerging and decaying patterns, like works in [25] [26], it cannot predict the likely occurrence times of emerging and decaying patterns in time-stamped datasets for decision making. For instance, in target-oriented decision support systems such as those for tracking fraudulent transactions, though techniques in [24] [25] [26] will be able to predict the continuous emergence of a fraudulent transactions, they are unable to predict the time-frame such transactions will occur. The ability to predict the likely occurrence times of emerging and decaying patterns can be useful in decision making such as, curbing crime, disease outbreak control, target oriented advertisement, and preventing customer attrition.

It is however worth noting that emerging pattern mining is quite different from incremental data mining techniques such as [14] [38] in the sense that most emerging pattern mining techniques consider static data while incremental data mining techniques deal with data streams. That notwithstanding, some works such as [12] mine emerging patterns in data streams.

Table 1. Sample Customer Transactions

TID	Transaction
T_1	$\{a, b, c, e\}$
T_2	$\{d, e, f\}$
T_3	$\{a, b, c, d\}$
T_4	$\{c, d, e, f\}$
T_5	$\{a, b, c, e, f\}$
T_6	$\{a, d, e, f\}$
T_7	$\{c, d\}$
T_8	$\{e, f\}$

To enable the prediction of the likely occurrence times of emerging and decaying patterns in time-stamped datasets for decision making, this work incorporates the concept of periodicity (from periodic frequent pattern mining) on EDTrend proposed in [24].

3. Definitions.

This section presents the formal notations and definitions in relation to frequent patterns, periodic frequent patterns as well as emerging and decaying patterns.

3.1. Frequent Patterns.

The problem of frequent pattern (itemset) mining is as follow. Let $I = \{i_1, i_2,..., i_m\}$ be a set of literals, called items. A set $X_1 = \{i_a, \ldots, i_n\} \subseteq I$, where $a \leq n$ and $a, n \in [1, m]$, is called a pattern (or an itemset). A transaction database is a set of transactions $D = \{T_1, T_2, T_3, \ldots, T_k\}$ such that for each transaction T_a, $T_a \in I$ and T_a has a unique identifier a called its transaction ID (TID). For example, consider the transaction database in Table 1 (a sample customer transaction database - which will be used the running example), the set of items for this database is $I = \{a, b, c, d, e, f\}$. Transaction T_2 in Table 1 which has a transaction ID of 2 and three items $\{d, e, f\}$ is a length-3 itemset.

The *coverset* of an itemset, S in a database, D, denoted as $cov(S)$ is defined as $cov(S) = \{m|m \in D \wedge S \subseteq m\}$. For example, in Table 1 given, $S = \{a, b\}$, then $cov(S) = \{1, 3, 5\}$ since $\{a, b\}$ appears in transactions 1, 3 and 5. The *support count* of S in D is defined as $|cov(S)|$ and the *support* of S in D, denoted as $sup(S)$ is defined as:

$$sup(S) = \frac{|cov(S)|}{|D|} \tag{1}$$

For instance, in Table 1, given $S = \{a, b\}$, then $sup(S) = \frac{3}{8} = 0.375$ as $|cov(S)| = |\{1, 3, 5\}| = 3$ and $|D| = 8$.

Definition 3.1. (*Frequent itemset mining* [2]). *The problem of frequent itemset mining consists of discovering frequent itemsets* [2]. *An itemset S is a frequent itemset in a database D if its support, $sup(S)$, is not less than a user-specified minimum support threshold,* minsup.

For example, considering a *minsup* threshold of 0.5 on Table 1, the set of frequent itemsets and their respective supports in Table 1 will be $\{a\} : 0.5, \{c\} : 0.625, \{d\} : 0.625, \{e\} : 0.75, \{f\} : 0.625$, and $\{e, f\} : 0.625$.

3.2. Periodic Frequent Patterns.

The formal definitions and notations commonly employed in periodic frequent pattern mining are presented as follows.

Definition 3.2. (*Consecutive transactions of an itemset*). *Let $D = \{T_1, T_2, T_3, \ldots, T_u\}$ be a database with u transactions. Let the set of transactions in D containing* *an itemset S be denoted as $g(S) = \{T_{g_1}, T_{g_2}, T_{g_3}, \ldots, T_{g_{n-1}}, T_{g_n}\}$, where, $1 \leq g_1 < g_2 < \ldots g_{n-1} < g_n \leq u$. Two transactions $T_x \supset S$ and $T_y \supset S$ are said to be consecutive with respect to S if there does not exist a transaction $T_w \in g(S)$ such that $x < w < y$. The period of two consecutive transactions T_x and T_y in $g(S)$ is defined as $p(T_x, T_y) = (y - x)$, that is the number of transactions between T_x and T_y.*

For example, consider the itemset $\{c\}$ in Table 1 which appears in transactions T_1, T_3, T_4, T_5, and T_7, then transactions, T_1 and T_3, or T_3 and T_4, or T_4 and T_5 or T_5 and T_7 are its consecutive transactions. The period between consecutive transactions T_1 and T_3 thus become $p(T_1, T_3) = 3 - 1 = 2$.

Definition 3.3. (*Set of all periods of an itemset*). *Let the set of transactions in a database D (with u transactions) containing an itemset S be denoted as $g(S) = \{T_{g_1}, T_{g_2}, T_{g_3}, \ldots, T_{g_{n-1}}, T_{g_n}\}$, such that, $1 \leq g_1 < g_2 < \ldots g_{n-1} < g_n \leq u$. The coverset of S in D become, $cov(S) = \{g_1, g_2, g_3, \ldots, g_{n-1}, g_n\}$. The set of all periods of S in D denoted as $ps(S)$ is defined as $ps(S) = \{g_1 - g_0, g_2 - g_1, g_3 - g_2, \ldots, g_n - g_{n-1}, |D| - g_n\}$, where $g_0 = 0$ is a constant and $|D|$ is the size of the database.*

For example, consider the itemset $\{a, b\}$ in Table 1 (where $|D| = 8$) which appears in transactions T_1, T_3, and T_5. The coverset of $\{a, b\}$ thus become, $cov\{a, b\} = \{1, 3, 5\}$, hence, the set of all periods of $\{a, b\}$ in D based on Definition 3.3 will be evaluated as $ps(\{a, b\}) = \{1 - 0, 3 - 1, 5 - 3, 8 - 5\} = \{1, 2, 2, 3\}$.

To mine the set of patterns with similar periodic occurrence times (shapes) in transaction databases for decision making, Nofong [27] define a periodic frequent pattern as follows.

Definition 3.4. (*Periodic frequent pattern* [27]). *Given a database $\mathbf{D}$, minimum support threshold ε, periodicity threshold p, difference factor p_1, a pattern S and $ps(S)$, S is a periodic frequent pattern if $sup(S) \geq \varepsilon$, $(p - p_1) \leq Prd(S) - std(ps(S))$ and $Prd(S) + std(ps(S)) \leq (p + p_1)$.*

where, $Prd(S)$ is the mean of $ps(S)$ - that is, $\bar{x}(ps(S))$ - is the periodicity of S and $std(ps(S))$ the standard deviation in $ps(S)$.

The periodicity in Definition 3.4 is what this paper incorporates in predicting the likely occurrence times of emerging and decaying patterns in time-stamped datasets.

3.3. Emerging Patterns.

Given two datasets, $\mathbf{D_i}$ and $\mathbf{D_{i+1}}$, the growth rate of a pattern S, $GR(S)$, from $\mathbf{D_i}$ to $\mathbf{D_{i+1}}$ is defined as [3]:

$$GR(S) = \begin{cases} 0, & \text{if } sup_{\mathbf{D_i}}(S){=}0 \wedge sup_{\mathbf{D_{i+1}}}(S) = 0 \\ \infty, & \text{if } sup_{\mathbf{D_i}}(S){=}0 \wedge sup_{\mathbf{D_{i+1}}}(S) > 0 \\ \frac{sup_{\mathbf{D_{i+1}}}(S)}{sup_{\mathbf{D_i}}(S)}, & \text{Otherwise} \end{cases} \tag{2}$$

For example, given the pattern $\{a, b\}$ that is mined from two datasets D_1 and D_2 having supports of 0.45 and 0.6 in D_1 and D_2 respectively, the growth rate of $\{a, b\}$ will be evaluated as $\frac{0.6}{0.4} = 1.5$.

With the growth rate, Dong and Li [3] introduced the concept to emerging pattern mining and define an emerging pattern as follows.

Definition 3.5. (Emerging pattern [3]). *Given $\rho > 1.0$ as the growth-rate threshold, a pattern S is said to be a ρ-emerging pattern (ρ-EP or simply EP) from $\mathbf{D_i}$ to $\mathbf{D_{i+1}}$ if $GR(S) \geq \rho$.*

Over the past years, it was observed that finding all emerging patterns above a minimum growth rate constraint as proposed in [3] often generates too many emerging patterns to be analysed. Various techniques were thus proposed for efficient mining of interesting categories of emerging patterns in works such as: jumping emerging patterns [8] [16] [34], essential emerging patterns [6] and interesting emerging patterns [7].

Though these proposed categories of emerging patterns have been useful in static datasets for classification, [25] [26] observed that in time-stamped datasets such emerging patterns often: are spikes or noise; contain some redundant emerging patterns; miss some useful emerging patterns; or, report some emerging patterns due to random chance - which could make them unsuitable for trend prediction.

To mine emerging patterns in time-stamped datasets suitable for trend prediction, Nofong *et al.* [26] redefined an emerging pattern (EP) as follows.

Definition 3.6. (Emerging pattern [26]**).** *Given ε as the minimum support, a pattern S is an emerging pattern from $\mathbf{D_i}$ to $\mathbf{D_{i+1}}$ if it is frequent in both $\mathbf{D_i}$ and $\mathbf{D_{i+1}}$ and $GR(S) > 1.0$.*

With Definition 3.6 and the concept of generator patterns, Nofong *et al.* [26] introduced the *non-derivable* emerging patterns as those without redundant information. The non-derivable emerging patterns was proposed in [26] to eliminate the redundant emerging patterns often reported in some existing works and to ensure reported emerging patterns are suitable for trend prediction. Nofong *et al.* [26] argued that redundant emerging patterns can be trivial if not detrimental in trend prediction. They subsequently showed (using naive techniques) the potential use and effectiveness of the set of non-derivable emerging patterns in trend prediction.

The work in [25] also observed that some reported emerging patterns might be emerging due to random chance without inherent item relationships. Nofong [25] argued that reporting patterns whose emergence are due to random chance without inherent item relations could be detrimental in trend prediction as patterns that are emerging due to random chance are often spiky, noisy or false trends. As such the set of *productive* emerging patterns (emerging patterns whose emergence are due to inherent item relations and not by random chance or data fluctuations) were introduced in [25].

For trend prediction, the works in [25] [26] employed naive approaches based on their introduced non-derivable and productive emerging patterns for trend prediction. Though both productive emerging and decaying patterns where used in [25] for trend prediction (unlike [26] which employed only non-derivable emerging patterns), no formal definition for decaying patterns was mentioned in [25].

Nofong in [24] proposed EDTrend, a methodology for trend prediction with emerging and decaying patterns. Adopting the definitions of both non-derivable and productive emerging patterns, non-derivable and productive decaying patterns were formally defined in [24]. Based on the defined categories of emerging and decaying patterns (non-derivable and productive), EDTrend was shown to be effective in predicting the continuous emergence and decay of patterns, as well as their likely supports.

4. **Incorporating Periodicity in Occurrence Time Prediction.** In time-stamped datasets such as those containing information about crimes of a city, customers transactions, or fraudulent credit card transactions, EDTrend ([24]) will be able to predict the continuous emergence or decay of crimes, customers transactions, or fraudulent credit card transactions for decision making. However, in targeted decision making (such as selective advertisements to a group of customers), predicting the continuous emergence

or decay of trends with EDTrend will not be sufficient enough as decision makers will be more interested in the possible occurrence times of trends.

In this work, EDTrend is enriched to predict not just the continuous emergence or decay of patterns, but their likely occurrence times for targeted decision making. To achieve this, the study herein proposes to incorporate a measure of periodicity in [27] in EDTrend for this prediction.

To predict the occurrence time of an emerging or decaying pattern with time, for a given minimum support ε, let S be an emerging or decaying pattern detected from $\mathbf{D_i}$ to $\mathbf{D_{i+1}}$ whose occurrence times in $\mathbf{D_{i+2}}$ are intended to be predicted. Let $cov_{\mathbf{D_i}}(S)$ and $cov_{\mathbf{D_{i+1}}}(S)$ be the coversets of S in $\mathbf{D_i}$ and $\mathbf{D_{i+1}}$ respectively. We can thus obtain $ps(S)_{\mathbf{D_i}}$ and $ps(S)_{\mathbf{D_{i+1}}}$ from $cov_{\mathbf{D_i}}(S)$ and $cov_{\mathbf{D_{i+1}}}(S)$. With $ps(S)_{\mathbf{D_i}}$ and $ps(S)_{\mathbf{D_{i+1}}}$, the periodic intervals of S and the standard deviations in its periods in both $\mathbf{D_i}$ and $\mathbf{D_{i+1}}$ can be evaluated, that is, $Prd_{\mathbf{D_i}}(S), std_{\mathbf{D_i}}(S), Prd_{\mathbf{D_{i+1}}}(S)$ and $std_{\mathbf{D_{i+1}}}(S)$ can be evaluated from $ps(S)_{\mathbf{D_i}}$ and $ps(S)_{\mathbf{D_{i+1}}}$.

With the evaluated periods and standard deviations in the periods of S in $\mathbf{D_i}$ and $\mathbf{D_{i+1}}$, we predict the occurrence times of S in $\mathbf{D_{i+2}}$ by treating the sets (Equations 3 and 4) as time-series:

$$T = \{Prd_{\mathbf{D_i}}(S), Prd_{\mathbf{D_{i+1}}}(S)\} \tag{3}$$

$$T_1 = \{std_{\mathbf{D_i}}(S), std_{\mathbf{D_{i+1}}}(S)\} \tag{4}$$

As such, similar approach employed in modelling EDTrend as discussed in [24] can be employed in estimating the expected occurrence times of S in $\mathbf{D_{i+2}}$.

For the occurrence time predictions, let $ePrd_{\mathbf{D_{i+2}}}(S)$ and $estd_{\mathbf{D_{i+2}}}(S)$ be the estimated periodicity and standard deviations of S in $\mathbf{D_{i+2}}$ obtained from $T = \{Prd_{\mathbf{D_i}}(S), Prd_{\mathbf{D_{i+1}}}(S)\}$ and $T_1 = \{std_{\mathbf{D_i}}(S), std_{\mathbf{D_{i+1}}}(S)$ respectively using the EDTrend algorithm. Based on $ePrd_{\mathbf{D_{i+2}}}(S)$ and $estd_{\mathbf{D_{i+2}}}(S)$, we predict the likely occurrence times of S in $\mathbf{D_{i+2}}$ as within the ranges:

$$\begin{aligned}
&1^{st} \text{ occurrence: within } t_1 \text{ to } t_2 \\
\text{where } t_1 = ePrd_{\mathbf{D_{i+2}}}(S) - estd_{\mathbf{D_{i+2}}}(S) &\text{ and } t_2 = ePrd_{\mathbf{D_{i+2}}}(S) + estd_{\mathbf{D_{i+2}}}(S) \\
2^{nd} \text{ occurrence: within } t_2 \text{ to } t_3, &\text{ where } t_3 = (t_2 + ePrd_{\mathbf{D_{i+2}}}(S)) \\
3^{rd} \text{ occurrence: within } t_3 \text{ to } t_4, &\text{ where } t_4 = (t_3 + ePrd_{\mathbf{D_{i+2}}}(S)) \\
4^{th} \text{ occurrence: within } t_4 \text{ to } t_5, &\text{ where } t_5 = (t_4 + ePrd_{\mathbf{D_{i+2}}}(S)) \\
&\text{and so on}
\end{aligned}$$

Though the proposed approach does not predict the actual occurrence time of S in $\mathbf{D_{i+2}}$ as a single value, the predicted ranges narrow down to a search window where the targeted decision making such identifying fraudulent card transactions, or selective advertisements will be more effective.

For a pattern S, the prediction of its occurrence times in $\mathbf{D_{i+2}}$ is precise if:

$$\sigma_1(ePrd_{\mathbf{D_{i+2}}}(S), Prd_{\mathbf{D_{i+2}}}(S)) < \Delta \tag{5}$$

$$\sigma_2(estd_{\mathbf{D_{i+2}}}(S), std_{\mathbf{D_{i+2}}}(S)) < \Delta \tag{6}$$

where, σ_1 and σ_2 are the deviations in: predicted and actual periods ($ePrd_{\mathbf{D_{i+2}}}(S)$ and $Prd_{\mathbf{D_{i+2}}}(S)$), and predicted and actual standard deviations ($estd_{\mathbf{D_{i+2}}}(S)$ and $std_{\mathbf{D_{i+2}}}(S)$) respectively, while Δ is a user controlled deviation threshold value.

For a given deviation threshold Δ, occurrence time predictions which have either one or both of σ_1 and σ_2 greater than Δ are identified as imprecise predictions.

5. **Experimental Results.** Experimental analysis was conducted on the following real-world databases shown in Table 2:

Table 2. Databases and Characteristics

Database	Characteristics	Time Periods	Source
Tafeng Retail	817741 transactions from 32266 unique customers	November 2000 to February 2001	AIIA Lab
Twitter	Hashtags and URLs extracted from tweets	1st November to 30th November 2012	CNetS

The aim of the predictions in the Tafeng Retail database is to use the monthly emerging and decaying customer transactions in predicting the likely future occurrence times of customer transactions while in the Twitter database, the daily emerging and decaying hashtags are used in predicting the likely occurrence times of hashtags in tweets for the next day.

The prediction precisions in each dataset are as discussed below. The following are the meaning of the terms in Tables 3, 4, 5, 6, 7 and 8

1. ε: The minimum support employed in mining the emerging and decaying patterns
2. P_{EP}: Prediction precision with only emerging patterns
3. P_{DP}: Prediction precision with only decaying patterns
4. P_{CB}: Prediction precision with combined emerging and decaying patterns.

5.1. **Twitter Dataset.** For occurrence time prediction in this dataset with the proposed periodicity measure, as can be seen in Tables 3, 4 and 5, prediction precisions decreases with decreasing minimum support. Also, it was observed and as can be seen in Tables 3, 4 and 5 that, the occurrence time predictions are more precise when Δ is set high. As can be observed in Tables 3, 4 and 5, the prediction precisions with only emerging patterns are relatively higher than the prediction precisions with only decaying patterns. It is also worth noting that the prediction precisions are higher when the productive emerging patterns are employed in occurrence time prediction compared to the non-derivable emerging patterns. This is because some non-derivable emerging and decaying patterns could be emerging or decaying due to random chance unlike the productive emerging and decaying patterns.

Table 3. Occurrence Time Prediction: Twitter Dataset at $\Delta = 8.0$

Days	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
1^{st}, 2^{nd} and 3^{rd}	3.0	**100.0**	50.0	88.9	**100.0**	50.0	90.0
	2.0	83.3	33.3	73.3	**84.6**	33.3	75.0

5.2. **Tafeng Retail Dataset.** For occurrence time prediction in this dataset with the proposed periodicity measure, similar observations as in the Twitter dataset can be made. That is, as can be seen in Tables 6, 7 and 8, prediction precisions decreases with decreasing minimum support. Similarly, setting Δ high results in higher prediction precisions as can be seen in Tables 6, 7 and 8.

Table 4. Occurrence Time Prediction: Twitter Dataset at $\Delta = 2.0$

Days	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
1^{st}, 2^{nd} and 3^{rd}	3.0	85.7	0.0	66.7	**87.5**	0.0	70.0
	2.0	50.0	0.0	40.0	**53.8**	0.0	43.8

Table 5. Occurrence Time Prediction: Twitter Dataset at $\Delta = 1.0$

Days	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
1^{st}, 2^{nd} and 3^{rd}	3.0	85.7	0.0	66.7	**87.5**	0.0	70.0
	2.0	50.0	0.0	40.0	**53.8**	0.0	43.8

Table 6. Occurrence Time Prediction: Tafeng Dataset at $\Delta = 8.0$

Months	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
Nov, Dec and Jan	7.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	6.0	**100.0**	76.9	84.2	**100.0**	76.9	84.2

Table 7. Occurrence Time Prediction: Tafeng Dataset at $\Delta = 4.0$

Months	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
Nov, Dec and Jan	7.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	6.0	**100.0**	76.9	84.2	**100.0**	76.9	84.2

Table 8. Occurrence Time Prediction: Tafeng Dataset at $\Delta = 2.0$

Months	ε (%)	Non-derivable			Productive		
		P_{EP}	P_{DP}	P_{CB}	P_{EP}	P_{DP}	P_{CB}
Nov, Dec and Jan	7.0	**100.0**	80.0	90.0	**100.0**	80.0	90.0
	6.0	**66.7**	53.8	57.9	**66.7**	53.8	57.9

6. **Conclusion.** This work has shown that incorporating periodicity (from periodic frequent pattern mining) in trend prediction with emerging and decaying patterns can enable accurate prediction of the likely occurrence times of emerging and decaying patterns. The ability to predict the likely occurrence times of emerging and decaying patterns can be useful in decision making such as selective advertisements to customers, identifying fraudulent transactions, crime control and so on. Future works on the applications of emerging and decaying patterns in trend prediction will look into extensions to enable predict the actual occurrence times of patterns based on its periodicity and not a range of intervals.

REFERENCES

[1] M. K. Afriyie, V. M. Nofong, J. Wondoh, and H. Abdel-Fatao, "Mining Non-redundant Periodic Frequent Patterns," *Asian Conference on Intelligent Information and Database Systems*, pp. 321–331, 2020.

[2] R. Agrawal, T. Imieliński, and A. & Swami, "Mining Association Rules between Sets of Items in Large Databases," *ACM SIGMOD Record*, Vol. 22(2), pp. 207–216, 1993.

[3] G. Dong and J. Li, "Efficient mining of emerging patterns: discovering trends and differences," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 43–52, 1999.

[4] G. Dong and J. Li, "Mining border descriptions of emerging patterns from dataset pairs," *Knowledge and Information Systems*, Vol. 8(2), pp. 178–202, 2005.
[5] G. Dong, X. Zhang, L. Wong, and J. Li, "CAEP: Classification by aggregating emerging patterns," *International Conference on Discovery Science*, pp. 30–42, 1999.
[6] H. Fan and K. Ramamohanarao, "An efficient single-scan algorithm for mining essential jumping," *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp.456–462, 2002.
[7] H. Fan and K. Ramamohanarao, "Efficiently mining interesting emerging patterns," *International Conference on Web-Age Information Management*, pp. 189–201, 2003.
[8] H. Fan and K. Ramamohanarao, "Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18(6), pp. 721–737, 2006.
[9] P. Fournier-Viger, C. W. Lin, Q. H. Duong, and T. L. Dam, L. Ševčík, D. Uhrin, and M. Voznak, "PFPM: discovering periodic frequent patterns with novel periodicity measures," *The Czech-China Scientific Conference*, pp. 27–38, 2017.
[10] P. Fournier-Viger, Z. Li, J. C. W. Lin, R. U. Kiran, and H. Fujita, "Discovering periodic patterns common to multiple sequences," *Big Data Analytics and Knowledge Discovery*, pp. 231–246, 2018.
[11] P. Fournier-Viger, P. Yang, J. C .W. Lin, and R. U. Kiran, "Discovering stable periodic-frequent patterns in transactional data," *The International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 230–244, 2019.
[12] Á. M. García-Vico, C. J. Carmona, P. González, H. Seker, and M. J. d. Jesus, "FEPDS: a proposal for the extraction of fuzzy emerging patterns in data streams," *IEEE Transactions on Fuzzy Systems*, Vol. 28(12), pp. 3193–3203, 2020.
[13] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, Vol. 29(2), pp. 1–12. 2000.
[14] T. P. Hong, C. Y. Wang, and Y. H. Tao, "A new incremental data mining algorithm using pre-large itemsets," *Intelligent Data Analysis*, Vol. 5(2), pp. 111–129, 2001.
[15] R. U. Kiran, A. Anirudh, C. Saideep, M. Toyoda, P. K. Reddy, and M. Kitsuregawa, "Finding periodic-frequent patterns in temporal databases using periodic summaries," *Data Science and Pattern Recognition*, Vol. 3(2), pp. 24–46, 2019.
[16] J. Li, G. Dong, and K. Ramamohanarao, "Making use of the most expressive jumping emerging patterns for classification," *Knowledge and Information Systems*, Vol. 3(2), pp. 1–29, 2001.
[17] J. Li, G. Dong, K. Ramamohanarao, and L. Wong, "Deeps: a new instance-based lazy discovery and classification system," *Machine Learning*, Vol. 54(2), pp. 99–124, 2004.
[18] J. Li, H. Liu, J. R. Downing, A. E. J. Yeoh, and L. Wong, "Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients," *Bioinformatics*, Vol. 19(1), pp. 71–78, 2003.
[19] J. Li and L. Wong, "Emerging patterns and gene expression data," *Genome Informatics*, Vol. 12, pp. 3–13, 2001.
[20] J. Li, K. Ramamohanarao, and G. Dong, "Combining the strength of pattern frequency and distance for classification," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 455–466, 2001.
[21] J. Lou and A. Cheng, "Detecting pattern changes in individual travel behavior from vehicle GPS/GNSS data," *Sensors*, Vol. 20(8), 2295, 2020.
[22] F. Neves, A. Finamore, and R. Henriques, "Efficient discovery of emerging patterns in heterogeneous spatiotemporal data from mobile sensors," EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. 1–10, 2020.
[23] V. M. Nofong, J. Liu, and J. Li, "A study on the applications of emerging sequential patterns," *Australasian Database Conference*, pp. 62–73, 2014.
[24] V. M. Nofong, "EDTrend: a methodology for trend prediction with emerging and decaying patterns," *Australasian Computer Science Week Multiconference*, pp. 1–8, 2016.
[25] V. M. Nofong, "Mining productive emerging patterns and their application in trend prediction," *Australasian Data Mining Conference*, pp. 109–117, 2015.
[26] V. M. Nofong, J. Liu, and J. Li, "Efficient mining of non-derivable emerging patterns," *Australasian Database Conference*, pp. 244–256, 2015.
[27] V. M. Nofong, "Discovering productive periodic frequent patterns in transactional databases," *Annals of Data Science*, Vol. 3(3), pp. 235–249, 2016.
[28] V. M. Nofong and J. Wondoh, "Towards Fast and Memory Efficient Discovery of Periodic Frequent Patterns," *Journal of Information and Telecommunication*, Vol. 3(4), pp. 480–493, 2019.

[29] V. M. Nofong, “Fast and memory efficient mining of periodic frequent patterns,” *Modern Approaches for Intelligent Information and Database Systems*, pp. 223–232, 2018.

[30] G. Poezevara, B. Cuissart, and B. Crèmilleu, “Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs,” *Journal of Intelligent Information Systems*, Vol. 37(3), pp. 333–353, 2011.

[31] H. S. Song, S.H. Kim, and J. K Kim, “Mining the change of customer behavior in an internet shopping mall,” *Expert Systems with Applications*, Vol. 21(3), pp. 157–168, 2001.

[32] A. Soulet, B. Crèmilleux, and F. Rioult, “Condensed representation of emerging patterns,” *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 127–132, 2004.

[33] S. K. Tanbeer, C. F. Ahmed, B. S. Jeong, & Y. K. Lee “Discovering periodic-frequent patterns in transactional databases,” *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 242–253, 2009.

[34] P. Terlecki and K. Walczak, “Jumping emerging patterns with negation in transaction databases - classification and discovery,” *Information Sciences*, Vol. 177(24), pp. 5675–5690, 2007.

[35] C. Tsai and Y. C. Shieh, “A change detection method for sequential pattern,” *Decision Support Systems*, Vol. 46(2), pp. 501–511, 2009.

[36] M. J. Zaki, “Scalable algorithms for association mining,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12(3), pp. 372–390, 2000.

[37] M. J. Zaki, and K. Gouda, K “Fast vertical mining using diffsets,” *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 326–335, 2003.

[38] H. Zheng, P. Li, Q. Liu, J. Chen, G. Huang, J. Wu, Y. Xue, and J. He, “Dual incremental fuzzy schemes for frequent itemsets discovery in streaming numeric data,” *Information Sciences*, Vol. 514, pp.15–43, 2020.

Vincent M. Nofong is a Lecturer in the Computer Science and Engineering Department at the University of Mines and Technology, Tarkwa, Ghana. He holds a PhD in Computer and Information Sciences from the University of South Australia, Adelaide. His current research interests include data mining, pattern mining, classification and trend prediction.

Hamidu Abdel-Fatao is a Lecturer in the Computer Science and Engineering Department at the University of Mines and Technology, Tarkwa, Ghana. He holds a PhD in Computer and Information Sciences from the University of South Australia, Adelaide. His research interests include GPS trajectory data mining, artificial intelligence, robotics and embedded computing, software and web development, bioinformatics.