A Sliding Windows based Dual Support Framework for Discovering Emerging Trends from Temporal Data

M Sulaiman Khan[1,2]
Dr Frans Coenen[2]
Dr David Reid[1]
Reshma Patel[3]
Lawson Archer[3]

[1]Liverpool Hope University

[2]University of Liverpool

[3]Transglobal Express Ltd. Wirral, UK

Outline of the Presentation

- Association Rule Mining
  - Downward closure property
- Temporal Association Rule Mining
- Jumping and Emerging Patterns
- Issues in Discovering JEPs
- Sliding Windows
- Dual support mechanism
  - DSAT Algorithm
  - Evaluation
- Conclusion & Future Work

Association Rule Mining

- Data Mining Technique for finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.
- Example: Customer buying Patterns from large market basket data/Transactions.
- Association rules are expressions of the form

$$X \rightarrow Y$$

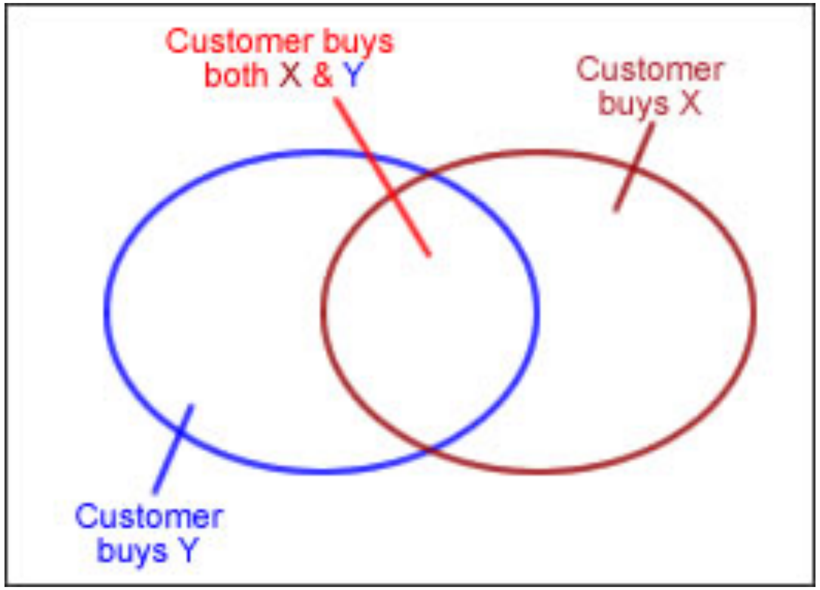
- where X and Y are item sets and $X \cap Y = \phi$

Interestingness Measures

Rule form: "Body → Head [support, confidence]".

We wish to find all rules of this form using the support confidence framework.

- Given a rule $X \ \& \ Y \Rightarrow Z$
  - **support**, ***s***, probability that a transaction contains {X & Y & Z}
  - **confidence**, ***c***, conditional probability that a transaction having {X & Y} also contains *Z*



Downward Closure Property

- Downward Closure Property (DCP)
  – Subsets of a frequent set are also frequent.
    e.g. if {A,B,C} is a frequent set then {A,B}, {A,C} and {B,C} will also be frequent.

  – Applications
    - Allows algorithms to efficiently generate frequent itemsets of increasing size by adding (K+1)-items to K-itemsets that are already ascertained to be frequent.

    - If itemsets {A,B} and {B,C} are not frequent, then (for example) {A,B,C} and {B,C,D} cannot be frequent, therefore there is no need to generate such “candidate” itemsets.

Temporal ARM (1)

- Temporal ARM (TARM) deals with the mining of time stamped databases, such as:
  - web server logs
  - super market transactional data
  - network traffic
- A TAR is an AR that exists during specific time intervals, for example:
  - flowers and chocolates are frequently sold together on the valentine day.
  - pumpkin and sweets are frequently sold together on Halloween.

Temporal ARM (2)

- Data mining technique directed at the identification of hidden trends in time series data
- In temporal ARM the attributes in the data are time stamped in some way as shown in table below:

Period	TID	Items	Period	TID	Items
January-09 (D1)	t_{01}	1,2,4	March-09 (D3)	t_{09}	4 6 8 10
	t_{02}	2,3		t_{10}	3 6 9
	t_{03}	1,2,3,4		t_{11}	1 3 4 7 8 9
	t_{04}	2,3,4		t_{12}	2 3 5 6 8 9
February-09 (D2)	t_{05}	1 3 5 7 9	April-09 (4)	t_{13}	4 9 10
	t_{06}	2 4 6 8 10		t_{14}	1 8 9
	t_{07}	1 2 4 5 7 8		t_{15}	2 3 5 7
	t_{08}	9		t_{16}	1

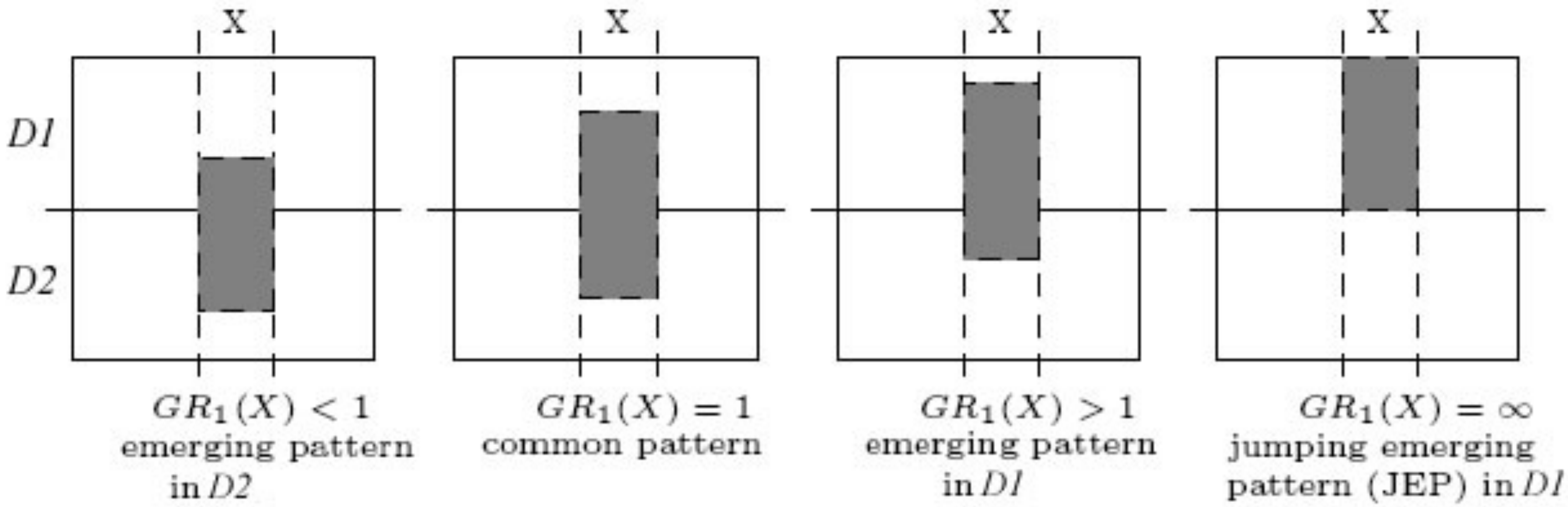
Jumping and Emerging Patterns

- One category of Temporal ARM is known as Jumping and Emerging Patterns (JEP) mining.
- An **Emerging Pattern** (EP) is usually defined as an itemset whose support increases over time according to some “change ratio” threshold.
- A **Jumping Pattern** (JP) is an itemset whose support changes much more rapidly than that for an EP.

Jumping Emerging Patterns

- Patterns whose frequency increases significantly from one data set to another

- Growth Rate of *X (patterns)* from D_2 to D_1



Growth Rate

$$GrowthRate(X) = \begin{cases} 0 & if\,(supp(X, D_1) = 0 \; and \; supp(X, D_2) = 0) \\ \infty & if\,(supp(X, D_1) = 0 \; and \; supp(X, D_2) \neq 0) \\ \dfrac{supp(X, D_2)}{supp(X, D_1)} & otherwise \end{cases}$$

$$GR(X) = \frac{supp(X, D_2)}{supp(X, D_1)} \longrightarrow GR(X) = \frac{supp(X, D_2)}{supp(X, D_1)} \times \frac{|D_1|}{|D_2|}$$

JEPs Example

Tid	Items	
T1 T2 T3 T4	A, B, C B, C, D, E B, C, E B, E	**D_1**
T5 T6 T7 T8	A, B, C, D A, B, C, D A, B, C A, D, E	**D_2**

- 2 datasets: D_1 & D_2
- 5 items: A, B, C, D, E
- *Supp*(ABC, D_1) = 1
- *Supp*(ABC, D_2) = 3
- *Supp*(BCD, D_1) = 1
- *Supp*(BCD, D_2) = 2

- *GR threshold = 2, JEPs from D_2 to D_1*
  - *ABC* is an emerging pattern (*GR*(*ABC*)=*3*)
  - *BCD* is not an emerging pattern (*GR*(*BCD*)=2)
  - *ABCD* is a jumping emerging pattern (*GR*(*ABCD*)=*infinity*)

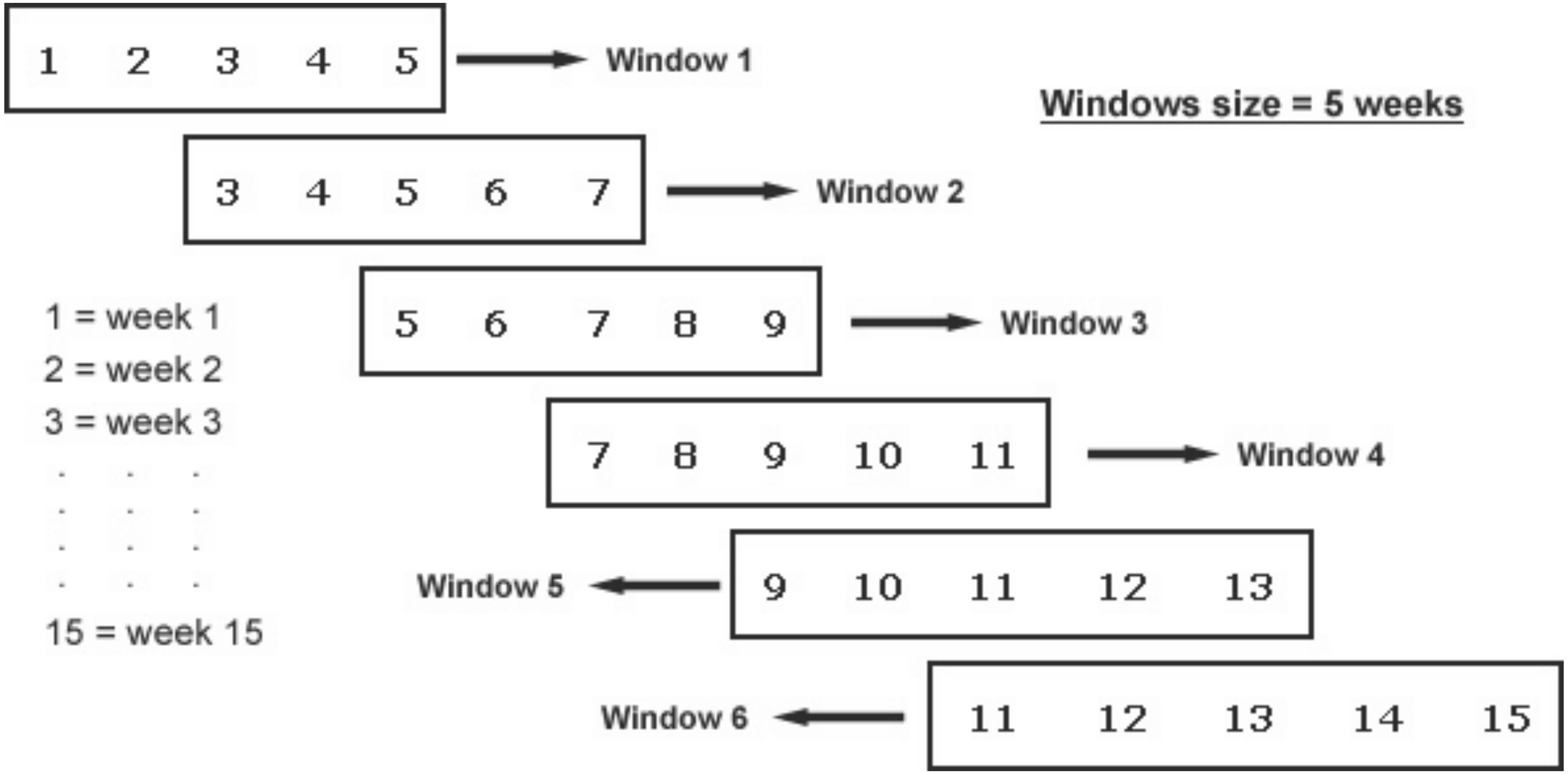
Issues in Discovering JEPs

- Discovering JEPs entails a significant computational overhead:
  - Large number of itemsets to compare (due to low threshold)
  - Data handling
  - Computational cost
  - Efficient memory management
- TARM processing models:
  - Landmark
  - Damped
  - Sliding Windows
- Maximal frequent set approach
  - Discovering of all JEPS is not guaranteed

Temporal ARM processing models

- Landmark Model
  - The Landmark model discovers all frequent itemsets over the entire history of data from a particular time called landmark to the current time.
- Damped Model
  - It is also known as Time-Fading model, finds frequent itemsets from temporal data in which each transaction is assigned a weight and this weight decreases with age. Older records contribute less weight toward itemset frequencies.
- Sliding Windows Model
  - The Sliding Windows model mines frequent itemsets in sliding windows. Only part of the transactions from a specific time period are stored in the sliding window and processed at the time when the window slides.

Sliding Windows Example



Dual Support Apriori for Temporal data (DSAT)

- Novel technique for discovering Jumping Emerging Patterns
- Mines time series data using a sliding window technique
- Utilizes the entire “data space” by avoiding itemsets borders with a constrained search space
- Avoids the computational overhead by exploiting previously mined time stamped data
- Discovers all JEPs, as in “naïve” approaches but utilises less memory and scales linearly with large datasets

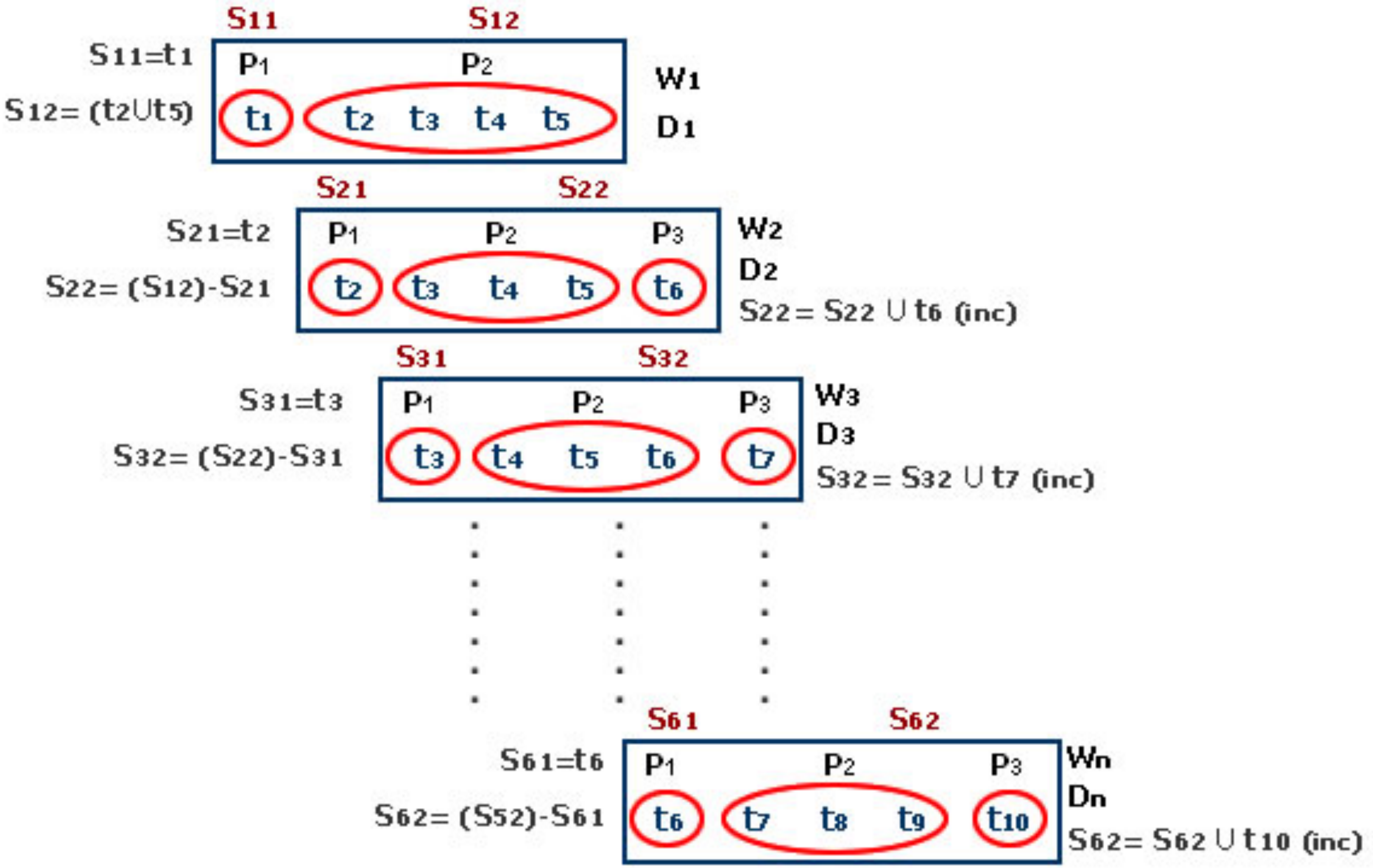
Dual support mechanism

- Each itemset holds two support counts called
  - *$Supp_1$*
  - *$supp_2$*

- ***$supp_1$*** *holds the support counts of itemsets in the "oldest" data segment that disappears whenever the window "slides"*

- ***$supp_2$*** *holds support counts for itemsets in the overlap between two windows and the recently added data segment.*

JEPs with dual support framework



DSAT Benefits

- The dual support mechanism utilises the already discovered frequent itemsets from the previous windows and avoids re-calculating support counts for all itemsets that exist in the overlapped datasets between two windows

- It only required databases access for the most resent segment, thus
  - less IO operations
  - less computation cost and
  - less memory utilization.

The DSAT Algorithm

- Dual Support Apriori Temporal (DSAT) algorithm comprises of two major steps:
  - Apply Apriori to produce a set of frequent itemsets using the sliding window approach.
  - Process and generate a set of JEPs such that the interestingness threshold (Growth Rate) is above some user specified threshold.

(Detail provided in paper)

Evaluation

- DSAT algorithm is evaluated with different datasets order to asses the
  - quality
  - efficiency and
  - effectiveness
- Datasets (server logs, point of sale, customer, synthetic)
  - Real and synthetic
  - Sparse and dense
  - Binary and quantitative

Experiments

- DSAT Performance
  - Comparisons with Apriori
  - Effect of varying data size
  - Effect of varying support threshold
  - Temporal effects of varying windows
  - Temporal effects of varying threshold
- Trend analysis example

Conclusions

- DSAT, a novel approach for
  - efficiently extracting JEPs
  - using sliding window
  - coupled with dual support mechanism
- Addressed issues in discovering JEPs
- Advantages of the framework:
  - less memory utilization
  - limited IO
  - fewer computations