# Enhanced Approximation of the Emerging Pattern Space using an Incremental Approach

(Preliminary report)

Walter M. Grandinetti[1] wmg@cs.uns.edu.ar

Carlos I. Chesñevar[2] cic@eps.udl.es

Marcelo A. Falappa[1] mfalappa@cs.uns.edu.ar

[1]Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur – Alem 1253 – B8000CPB Bahía Blanca, Argentina

[2]Grupo de Inteligencia Artificial – Departament d'Informàtica i Enginyeria Industrial
Universitat de Lleida – C/Jaume II, 69 – 25001 Lleida, Catalunya (Spain)

**Abstract**

From the many different patterns that can be extracted from data, so-called *emerging patterns* (EPs) are a particular useful kind. EPs are itemsets whose supports increase significantly from one dataset to another. Existing methods used to discover EPs have been successfully applied only under a constrained search space. Although they may provide a very efficient way of discovering some sort of EPs, they are rather limited when the whole set of EPs is needed, as they just compute an approximation of that set. Recent EPs techniques rely on *borders*, a concise representation of the candidate itemsets which does not require computing an exponentially large number of such candidates. In this paper we outline a new method which exploits previously mined data using an incremental approach, requiring thus less dataset accesses. Our proposal also aims to reduce the amount of work needed to perform difference operations among borders taking into account special properties of the itemsets.

**Key Words:** Pattern Mining, Emerging Patterns, Maximal Patterns, Incremental Mining.

## 1 Introduction and motivations

From the many different patterns that can be extracted from data, so-called *emerging patterns* (EPs) are a particular useful kind. EPs are itemsets whose supports increase significantly from one dataset to another. They are especially useful to point out changes and differences between datasets, and can also capture emerging trends when applied to timestamped databases. In particular, it has been shown that EPs with a low to medium support can give new insights and guidance to experts, in even "well understood" applications (Dong and Li 1999).

As an example consider a recent discovered trend[1], concerning the emerging trends of American students studying in Canadian Universities: the enrollments of American students in Canada have been rising by about 85% in three years to a total of about 5000. This trend is an emerging pattern (EP) with low support ($\frac{5000}{N}$, where $N$ is the total number of students enrolled in Canada) but a large growth rate (1.85). Previous methods used to discover this kind of information about the data have been successfully applied only under a constrained search space. Although they may provide a very efficient way of discovering some sort of EPs, they are rather limited when the whole set of EPs is needed, as they just compute an approximation of that set.

[1]Published in the newspaper *Dayton Daily News*, Ohio, USA, 10/6/2002.

In this paper we outline a new method which exploits previously mined data using an incremental approach, requiring thus less accesses to the dataset. Our proposal also aims to reduce the amount of work needed to perform the difference operations among borders taking into account some well-known properties of the itemsets.

## 2 Background

Emerging patterns (EPs) have been thoroughly investigated in recent years (Li 2001; Li and Dong 2004; Dong and Li 1999; Li and Wong 2002; Li et al. 2000; Bailey et al. 2002). Several classes of EPs can be distinguished, along with different proposal to approximate their associated emerging pattern space. The following definitions are extracted from (Dong and Li 1999). A set $X$ of items is called an *itemset*. A transaction $T$ contains an itemset $X$ if $X \subseteq T$. The *support* of $X$ in a dataset $\mathcal{D}$ (denoted $supp_{\mathcal{D}}(X)$) is $\frac{count_{\mathcal{D}}(X)}{|\mathcal{D}|}$ where $count_{\mathcal{D}}(X)$ is the number (called *count*) of transactions in $\mathcal{D}$ containing $X$. Given a number $\sigma > 0$, an itemset $X$ is $\sigma$-*large* in $\mathcal{D}$ if $supp_{\mathcal{D}}(X) \geq \sigma$. The collection of all $\sigma$-*large* is denoted as $\textsc{Large}_{\sigma}(\mathcal{D})$. Conversely, the collection of all itemset $\sigma$-*small* ($supp_{\mathcal{D}}(X) < \sigma$) is denoted as $\textsc{Small}_{\sigma}(\mathcal{D})$. For sake of simplicity, we will just write $supp_i(X)$ to denote $supp_{\mathcal{D}_i}(X)$. The *growth rate* of an itemset $X$ from a dataset $D_1$ to $D_2$ is defined as

$$GrowthRate(X) = \begin{cases} 0 & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) = 0 \\ \infty & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) \neq 0 \\ \frac{supp_2(X)}{supp_1(X)} & \text{otherwise} \end{cases}$$

Given a growth rate threshold $\rho > 1$, an itemset $X$ is called a $\rho-$*emerging pattern* ($\rho-$EP or simply EP) from $\mathcal{D}_1$ to $\mathcal{D}_2$ if $GrowthRate(X) \geq \rho$. The *EP* mining problem for a given growth-rate threshold is to find all $\rho-$EP [Dong & Li, 1999].

The EP mining problem can be best pictured in a 2-D support plane (Fig. 1), where every point $(\sigma_2, \sigma_1)$ represents an itemset $X$ such that $(supp_2(X), supp_1(X))$. The point $G = (\theta_{min}, \delta_{min})$ is identified in order to distinguish the sets $\textsc{Large}_{\delta_{min}}(\mathcal{D}_1)$ and $\textsc{Large}_{\theta_{min}}(\mathcal{D}_2)$. Given a fixed growth rate threshold $\rho$, the supports of all $\rho$-EPs from $\mathcal{D}_1$ to $\mathcal{D}_2$ must fall on the triangle $\Delta ACE$.

In (Dong and Li 1999) the EP mining problem is decomposed in three different areas. They proposed an efficient way to mine one of those areas and revealed problems related to mine the other ones. In particular, they found a highly efficient way to mine the zone bounded by the $BCDG$ rectangle by using *borders*, a concise representation of the candidate itemsets which does not require computing an exponentially large number of such candidates. The usage of borders allows to detect EPs using a small fraction of itemsets that represent a large number of candidates (Fig. 1). According to (Li and Dong 2004) a large number of EPs fall in the region $\Delta ABG$ because there is a large quantity of patterns with low support in each dataset. Hence it is a challenge for current techniques to find EPs within this region.

In order to get a better approximation to the emerging pattern space, (Dong and Li 1999) propose an approach to find EPs in the $\Delta GDE$ region applying recursively the algorithm used in the $BCDG$ rectangle onto the region $\Delta GDE$, identifying a new $B'C'D'G'$ rectangle (fig. 2).

In a more recent paper (Li and Dong 2004), Li and Dong suggest another way to approximate the set of EPs for any fixed growth rate threshold $\rho$. by ussing a *sequence* of $BCDG$ rectangles, exploiting the relationship $\delta_{min} \times \rho = \theta_{min}$. This approach is based on mining several $BCDG$ rectangles exploiting thus the highly efficient method previously mentioned. However, in order to perform such mining different borders are required (two for every $BCDG$ rectangle). In

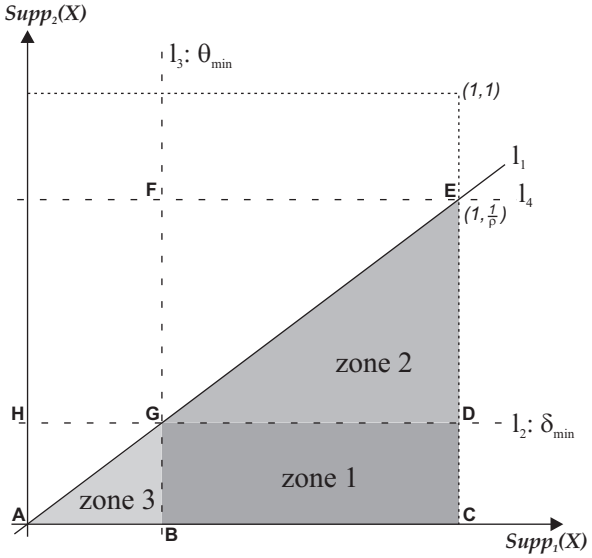


Figure 1: The support plane

order to do this, maximal patterns should be extracted from each dataset to form the borders for each different threshold. Clearly, such maximal pattern mining is computationally expensive and makes the approach quite inefficient.

Although both approaches can approximate better the emerging pattern space than the plain method proposed in (Li and Dong 2004) it can be pointed out that both suffer from several drawbacks. First, let us consider the recursive mining method shown in Fig. 2. There are two main drawbacks:

1. It does not take into account previously accomplished computation. It should be noted that the mining of the rectangle $BCDG$ takes as an input the maximal itemsets of both datasets. It seems to be likely that there is some relation between these maximal itemsets and the new ones to be computed. This relationship could improve the performance of the algorithm, but it is not exploited.

2. It is unclear how the projected database can be efficiently obtained. Probably, it is the result of intersecting both borders or maybe the whole database is used again.

Next, let us consider the alternative approach, which involves mining several $BCDG$ rectangles changing the threshold on either dataset. Although it does not have the problem of dealing with a subset of the dataset, it also has some drawbacks, namely:

1. It does not take into account previously accomplished computation. The reasons and consequences are analogous to the ones discussed for the previous approach.

2. Each rectangle overlaps the other rectangles. Thus, the final set of EPs should be recomposed taking into consideration that it is not a simple operation but a union of sets with many items duplicated among sets. The main problem is that neither the union nor the handling of duplicates are defined.

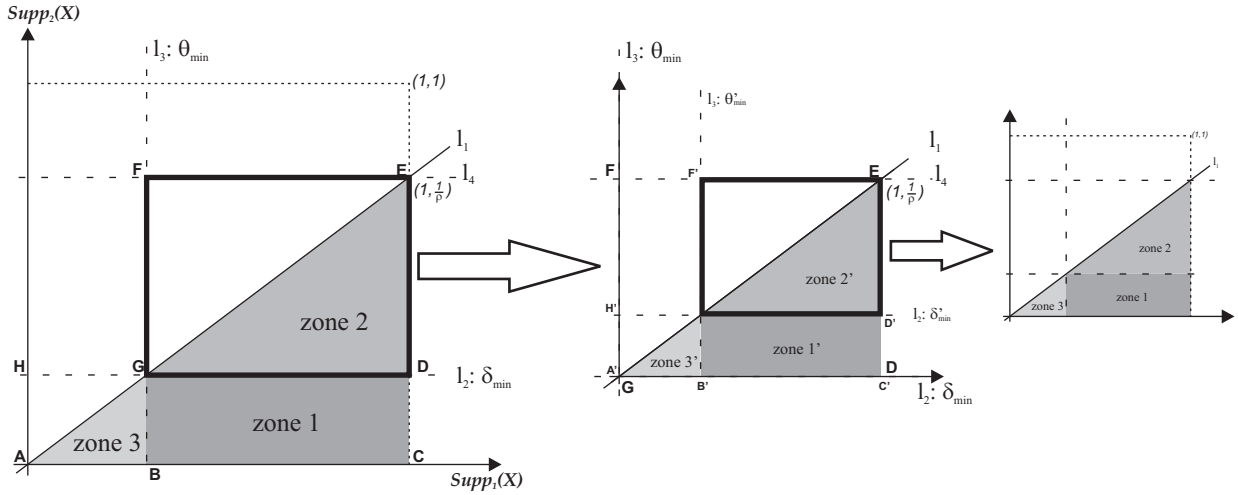


Figure 2: Result of mining the GDE triangle (Fig. 1)

## 3 From Riemman Sum to EP Mining

Based on the observations given in section 2, we propose a different way of exploring the EPs-support space. In the same way that Riemman Sum approximates the area enclosed below any function using the sum the areas of the adjacent rectangles (*sectors*), we propose to divide the $\Delta ACE$ triangle area into $k$ rectangles of equal width of length $\lambda$ (Fig. 3).

We aim to exploit previous known information in order to avoid repetitive computation. Hence, we propose to generate the collection of maximal patterns of the sector $i-1$ using the maximal patterns of the previous sector $i$ for any $1 \leq i \leq k$.

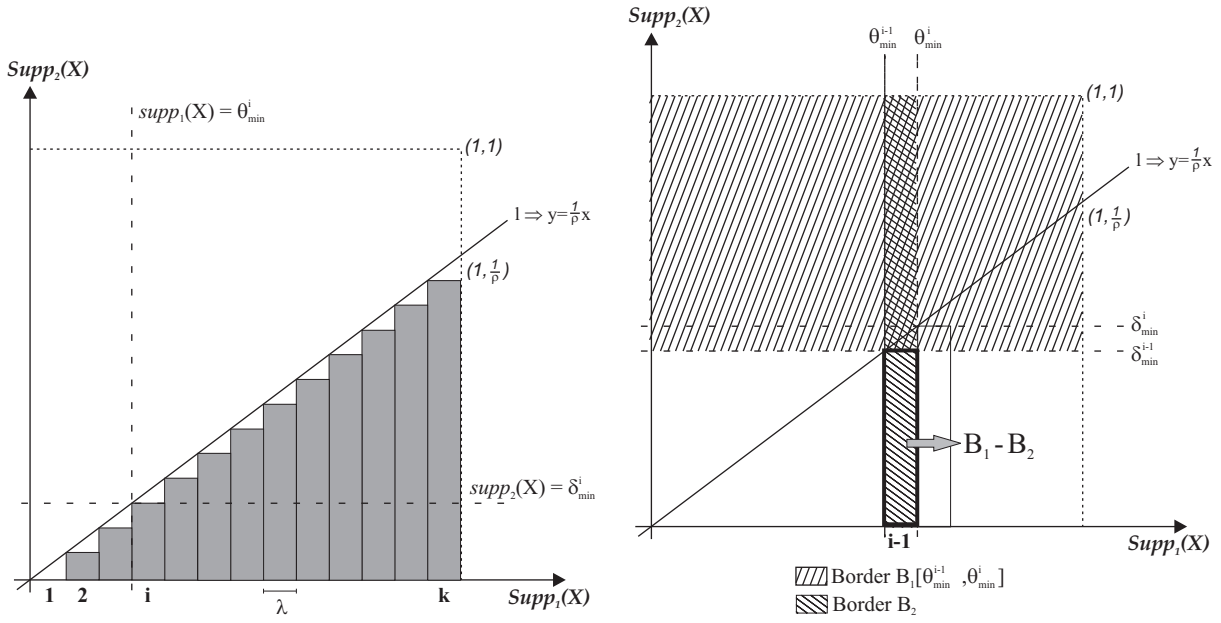


Figure 3: (a) Exploring the EPs support space using adjacent rectangles; (b)EPs Mining of sector $i-1$

The method proposed should begin obtaining those EPs that lays in sector $k$ using the basic algorithm. As a result, the emerging patterns of this sector will be discovered along with the

maximal patterns that were used to form the borders used to mine sector $k$. The mining task for a sector $i \in [1, k-1]$ will proceed from maximal patterns of previously mined sectors, in particular from sector $i+1$. Unlike sector $k$ (where a complete mining of the datasets is needed) the mining procedure for sector $i$ can take advantage of previous knowledge in order to improve the efficiency of the mining (either in time or in space). As it was already pointed out the EP-Miner needs the maximal patterns from both datasets in order to perform the mining. These maximal patterns can be either recomputed from scratch (as it is done in other techniques) or obtained by exploiting the way the mining is performed, namely updating the maximal patterns to fit the changes. Thus, the main idea of our approach is to produce the maximal patterns of sector $i-1$ from the maximal patterns of sector $i$.

The EPs mining of sector $i-1$ will proceed as follows:

1. Compute maximal patterns for $supp_1(X) = \theta_{min}^{i-1}$ and $supp_2(X) = \delta_{min}^{i-1}$.
2. Produce a border $B_1$ from maximal patterns in $\mathcal{D}_1$ of previous step.
3. Produce a border $B_1[\theta_{min}^{i-1}, \theta_{min}^{i}]$ for $\mathcal{D}_2$ where every itemset $X$ which belongs to this border satisfies $supp_1(X) \in [\theta_{min}^{i-1}, \theta_{min}^{i}]$ (Fig. 3). Then, the maximal patterns from sector $i$ are the left bound and the new maximal patterns are the right bound of the border, i.e. $B_1[\theta_{min}^{i-1}, \theta_{min}^{i}] = \langle \mathrm{MFI}_{\theta_{min}^{i}}, \mathrm{MFI}_{\theta_{min}^{i-1}} \rangle$.
4. As a last step, the difference operation is applied over the borders from the previous steps. Formally, $B_1[\theta_{min}^{i-1}, \theta_{min}^{i}] - B_2$.

## 4 Conclusions

In this paper we have summarized the main elements of a new approach for mining EPs, based on refining existing techniques using borders. Existing methods used to discover EPs have been successfully applied only under a constrained search space, being rather limited when the whole set of EPs is needed, as they just compute an approximation of that set. The proposed methodology exploits previously mined data using an incremental approach, requiring thus less dataset accesses. Our proposal also aims to reduce the amount of work needed to perform difference operations among borders taking into account special properties of the itemsets. Current research is focused on performing experiments in order to assess empirically how much computational effort is actually saved. An implementation of the proposed algorithm is underway.

## References


Bailey, J., T. Manoukian, and K. Ramamohanarao (2002). Fast algorithms for mining emerging patterns. In *PKDD*, pp. 39–50.

Dong, G. and J. Li (1999). Efficient mining of emerging patterns: Discovering trends and differences. *ACMKDD*.

Li, J. (2001, January). *Mining Emerging Patterns to Construct Accurate and Efficient Classifiers*. Ph. D. thesis, The University of Melbourne.

Li, J. and G. Dong (2004). Mining border description of emerging patterns from dataset pairs. Technical report, Wright University, USA.

Li, J., K. Ramamohanarao, and G. Dong (2000). The space of jumping emerging patterns and its incremental maintenance algorithms. In *ICML*, pp. 551–558.

Li, J. and L. Wong (2002). Geography of differences between two classes of data. In *PKDD*, pp. 325–337.