Random Sampling over Data Streams for Sequential Pattern Mining

Chedy Raïssi
LIRMM, EMA-LGI2P/Site EERIE
161 rue Ada
34392 Montpellier Cedex 5, France
France
raissi@lirmm.fr

Pascal Poncelet
EMA-LGI2P/Site EERIE
Parc Scientifique Georges Besse
30035 Nîmes Cedex, France
Pascal.Poncelet@ema.fr

February 16, 2007

Abstract

In recent years the emergence of new real-world applications such as network traffic monitoring, intrusion detection systems, sensor network data analysis, click stream mining and dynamic tracing of financial transactions, calls for studying a new kind of model. Named *data stream*, this model is in fact a continuous and potentially infinite flow of information as opposed to finite and statically stored data sets. We study the problem of sequential pattern mining in data streams. This problem has been extensively studied for the conventional case of disk resident data sets. In the case of data streams, this problem becomes more challenging as the volume of data is usually too huge to be stored on permanent devices, main memory or to be scanned thoroughly more than once. In this case, it may be acceptable to generate approximable solutions for our mining problem. In this paper we introduce a new approach based on biased reservoir sampling to achieve a more efficient mining of sequential patterns. Furthermore, we theoretically prove that our biased reservoir size is always bounded whatever the size of the stream is. This property often allows us to keep the entire relevant reservoir in main memory. We also show a simple algorithm to build the biased reservoir for the special case of sequential pattern mining. Experimental evaluation supports the claim that sequential pattern mining based on biased reservoir sampling needs small memory requirements. Besides, we also propose an adapted approach to handle the case of mining sequential patterns in a sliding window model. The experiment show that the results are accurate.

1 Introduction

Recently, the data mining community has focused on a new challenging model where data arrives sequentially in the form of continuous rapid streams. It is often referred to as data streams or streaming data. Since data streams are continuous, high-speed and unbounded flow of informations, it is often impossible to mine patterns with classical algorithms that require multiple scans. As a consequence new approaches were proposed to mine itemsets [5, 3, 2, 4, 8] using different approaches based on the *landmark*, *sliding windows* or *time-fading* models. However, few researches focused on sequential patterns extraction over data streams. In this paper we consider that transactions are ordered into the streams and grouped under different identifiers. We propose a new approach to mine sequential patterns based on the maintenance of a synopsis of the data stream. This proposition is motivated by the fact that the volume of data in real-world data streams is usually too huge to be efficiently mined and that an approximate answer for mining tasks is largely acceptable. In other words, in the data stream model one has to trade off accuracy against efficiency. A number of synopsis structures have been developed in recent years like sketches, sampling, wavelets and histograms. Our method belongs to the class of reservoir sampling. The reservoir sampling method is very easy to understand as it generates a sample of the original data representation. However, the classical unbiased reservoir method is inaccurate for data streams, this is due to the fact that when the the stream length increases, the accuracy of the reservoir decreases as it will contain a large portion of points from the distant history of the stream (the probability of successive insertions of new points reduces with the progression of the stream) and in an evolving data stream only the more recent data may be relevant for many mining tasks. To overcome this problem we use a biased reservoir sampling based on a temporal bias function in order to regulate the choice of the stream sample.

2 Preliminary Concepts

2.1 Sequential Patterns

The traditional sequence mining problem was first introduced in [7] and extended in [6].

Let $\mathcal{D}$ be a database of customer transactions where each transaction T consists of :

1. a customer-id, denoted by C_{id}
2. a transaction time, denoted by *time*
3. a set of items (called *itemset*) involved in the transaction, denoted by *it*

Definition 1 (Sequence) *Let $\mathcal{I} = \{i_1, i_2 ... i_m\}$ be a finite set of literals called items. An itemset is a non-empty set of items. A sequence S is a set of itemsets ordered according to their timestamp. It is denoted by $< it_1\ it_2\ ... it_n >$, where it_j, $j \in 1...n$, is an itemset. A k-sequence is a sequence of k items (or of length k).*

Definition 2 (Support) *Let C_{trans} be the ordered list of transactions for a single customer C (the maximal sequence supported by C). The support of a sequence S in a transaction database $\mathcal{D}$, denoted by $Support(S, \mathcal{D})$, is defined as:*

$$Support(S, \mathcal{D}) = \frac{|\{C \in \mathcal{D} | S \preceq C_{trans}\}|}{|\{C \in \mathcal{D}|}$$

Given a minimal support threshold, the problem of sequential pattern mining is to extract all the sequences S in $\mathcal{D}$ such that $Support(S, \mathcal{D}) \geq \sigma$.

2.2 Biased reservoir sampling

Reservoir sampling was first introduced in [9], in this method the first n points in the data stream are stored at the initialization step, when the $(t+1)^{th}$ is received in the data stream it replaces randomly one of the points in the reservoir with probability $\frac{n}{t+1}$. As the stream length increases the probability of the insertion reduces. This is a clear disadvantage for mining tasks that consider that recent information provided by the stream is the most relevant. One solution proposed in [1] was to use an exponential bias function defined as follow : $f(r, t) = e^{-\lambda(t-r)}$

with parameter λ being the bias rate. The aim of this bias function is to regulate the choice of the stream sample. In other words, the bias function modulates the sample in order to focus on recent or old behaviors in the stream following application specific constraints. Moreover, the inclusion of an exponential bias function implies also an upper bound on the reservoir size which is independant of the stream length. For a stream of length t, let $R(t)$ be the maximal size of the reservoir which satisfies the exponential bias function, we have $R(t) < \frac{1}{\lambda}$.

In the next section we give results on the bounds for sample size, given the desired accuracy of the results in terms of support and errore rate.

3 Sampling analysis

The first question that one has to answer when sampling for mining tasks is : how accurate my sample is compared to my original data set? We answer to this question by giving exact bounds on the size of the sample w.r.t an error rate.

Definition 3 (Error rate) *Let $\mathcal{D}$ be a database of customer transactions and $\mathcal{S}_\mathcal{D}$ a random sample generated from $\mathcal{D}$. Let s be a sequence from $\mathcal{D}$. The absolute error rate in terms of support estimation, denoted ϵ, is defined as :*

$$\epsilon(s, \mathcal{S}_\mathcal{D}) = |Support(s, \mathcal{S}_\mathcal{D}) - Support(s, \mathcal{D})|$$

Let X_i be a random variable for the i^{th} customer with $Pr[X_i = 1] = p_i$ if the i^{th} customer supports the sequence s and $Pr[X_i = 0] = 1 - p_i$, if not. All the X_i are independant. Note that we are in presence of Poisson trials as the number t of trials in which the probability of success p_i varies from trial to trial. Let $X(s, \mathcal{S}_\mathcal{D}) = \sum_i X_i = Support(s, \mathcal{S}_\mathcal{D}) \times |\mathcal{S}_\mathcal{D}|$ be the number of customers in the sample that supports the sequence s. Then the expected number of customers that support the sequence s in the sample is $E[X(s, \mathcal{S}_\mathcal{D})] = Support(s, \mathcal{D}) \times |\mathcal{S}_\mathcal{D}|$. We would like to estimate the probability that our error rate gets higher than a user defined threshold ϵ, denoted $Pr[e(s, \mathcal{S}_\mathcal{D}) > \epsilon]$.

Using Chernoff bounds the following theorem gives us a lower bound on the size of the reservoir given ϵ and a maximum probability δ that the error rate exceeds ϵ:

Theorem 1 *Given a sequence s then $Pr[e(s, \mathcal{S}_\mathcal{D}) > \epsilon] \leq \delta$ iff the reservoir size is $|\mathcal{S}_\mathcal{D}| \geq \ln(\frac{2}{\delta})\frac{1}{2\epsilon^2}$*

As we are working on biased reservoir samples, the following corollary gives an upper bound on the bias rate :

Corollary 1 *Given an error bound ϵ and a maximum probability δ that $e(s, \mathcal{S}_\mathcal{D}) > \epsilon$ we get an upper bound on the bias rate:*

$$\lambda \leq \frac{2\epsilon^2}{ln(2/\delta)}$$

4 Algorithm

Based on the sampling analysis results we built an algorithm that achieve exponential bias with the λ parameter. Unlike the algorithms presented in [1], our approach need to take into account the constraint of the lower bound of the size of the reservoir. Note that the reservoir size is defined in term of customers number and not in term of transaction numbers. That means that insertion and delete operations must be done at the customers level *and* at the itemsets level.

Algorithm 1: RESERVOIR SAMPLING FOR SEQUENTIAL PATTERNS algorithm

Data: Reservoir $\mathcal{S}_\mathcal{D}$;Bias rate λ;Transaction T
Result: Reservoir $\mathcal{S}_\mathcal{D}$ after insertion of the $(t+1)^{th}$ transaction

// $F(t) = \frac{q}{|\mathcal{S}_\mathcal{D}|} \in [0,1]$ is the fraction of the reservoir filled at the arrival of the t^{th} transaction
// $I(C_i, t) = \frac{i}{|itemsetList|} \in [0,1]$ is the fraction of the itemsets list for customer C_i at the arrival of the t^{th} transaction
if $T.C_i \notin \mathcal{S}_\mathcal{D}$ **then**
  // Deterministic insertion of the transaction T with its customer id C_i
  $Coin \leftarrow Random(0,1)$;
  if $Coin \leq F(t)$ **then**
    // Success case: we replace one of the customers with all its itemsets with T.
    $pos \leftarrow Random(0,q)$;
    Replace($T.C_{pos}.it$,$T.C_i.it$);
  else
    //Failure case: we directly add the transaction T without replacement.
    Add($T.it$,$\mathcal{S}_\mathcal{D}$);
    q++;
else
  // A sample of customer C_i transactions is already present in the reservoir
  // Deterministic insertion of the transaction T in C_i itemsets list
  $Coin \leftarrow Random(0,1)$;
  if $Coin \leq F(t)$ **then**
    $pos \leftarrow Random(0,i)$;
    ReplaceItemset(it_{pos},$T.C_i.it$);
  else
    AddItemset($T.it$,$C_i.itemsetList$);
    i++;

5 Experimentation

The experiments were performed on a Core-Duo 2.16 Ghz MacBook Pro with 1GB of main memory, running Mac OS X 10.4.6. We performed several tests with sythetic datasets that were generated with the IBM QUEST synthetic data generator, our data stream is divided into batches of period 25 seconds, each batch contains from 25k to 50k transactions. The memory management is the main focus of our performance study.

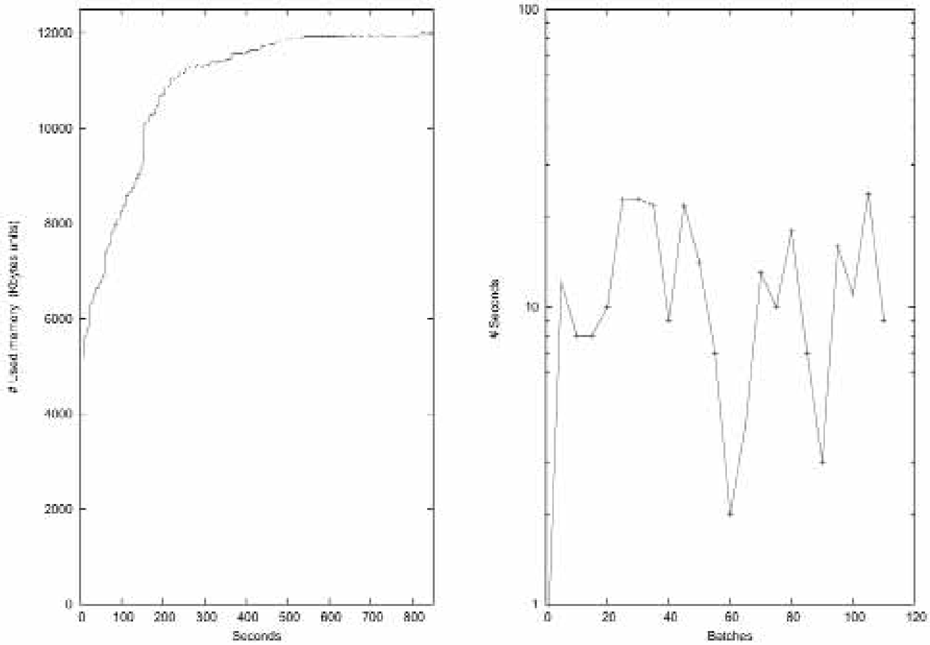

Figure 1: memory usage and time requirements for data set C1200I10K with reservoir sizes $\lambda = 2.10^{-5}$

6 Summary

In this presentation we introduced a new biased reservoir sampling algorithm for sequential pattern mining over data streams. The sampling analysis shows that we can efficiently bound our error rate to get approximate but acceptable results on our mining task. The experiments shows that our reservoir memory requirement are very low.

References

[1] Charu C. Aggarwal. On biased reservoir sampling in the presence of stream evolution. In Umeshwar Dayal, Kyu-Young Whang, David B. Lomet, Gustavo Alonso, Guy M. Lohman, Martin L. Kersten, Sang Kyun Cha, and Young-Kuk Kim, editors, *VLDB*, pages 607–618. ACM, 2006.

[2] Y. Chi, H. Wang, P.S. Yu, and R.R. Muntz. Moment: Maintaining closed frequent itemsets over a stream sliding window. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 04)*, pages 59–66, Brighton, UK, 2004.

[3] G. Giannella, J. Han, J. Pei, X. Yan, and P. Yu. Mining frequent patterns in data streams at multiple time granularities. In *In H. Kargupta, A. Joshi, K. Sivakumar and Y. Yesha (Eds.), Next Generation Data Mining, MIT Press*, 2003.

[4] H.-F. Li, S.Y. Lee, and M.-K. Shan. An efficient algorithm for mining frequent itemsets over the entire history of data streams. In *Proceedings of the 1st International Workshop on Knowledge Discovery in Data Streams*, Pisa, Italy, 2004.

[5] G. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB 02)*, pages 346–357, Hong Kong, China, 2002.

[6] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology (EDBT 96)*, pages 3–17, Avignon, France, 1996.

[7] R. Agrawal R. Srikant. Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering (ICDE 95)*, pages 3–14, Tapei, Taiwan, 1995.

[8] W.-G. Teng, M.-S. Chen, and P.S. Yu. A regression-based temporal patterns mining schema for data streams. In *Proceedings of the 29th International Conference on Very Large Data Bases (VLDB 03)*, pages 93–104, Berlin, Germany, 2003.

[9] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

Random Sampling over Data Streams for Sequential Pattern Mining

C. Raïssi et P. Poncelet

LIRMM, LGI2P/Ecole des Mines d'Alès

15 mars 2007

WDSA2007, Caserta, Italy.

Plan
Motivations
Preliminary Concepts
Related Work
Sequential Pattern Mining
Synopsis Construction
Sampling in static databases
Extending to Data Streams
Experimental Results
Conclusion and Summary

Motivations

- A new problem : data modeled as a potentially infinite flow of transactions
- Many recent real-world applications :
  1. Network traffic monitoring
  2. Trend analysis
  3. Sensor network data analysis

- Classical mining approaches are inefficient for this new problem
- In many cases, it may be acceptable to generate *approximate solutions* : synopsis structures ?

Definitions

- Let $\mathcal{D}$ be a database of customer transactions where each transaction T consists of :
  1. A customer-id, denoted by C_{id}
  2. A transaction time, denoted by *time*
  3. A set of items (called *itemset*) involved in the transaction, denoted by *it*

Example

Consider the following database $\mathcal{D}$ with $\mathcal{I} = \{a, b, c, d\}$:

C_1	T_1	a,b,c,d
	T_2	a,b
C_2	T_3	a,b
C_3	T_1	a,d
	T_4	c

Sequence

- Let $\mathcal{I} = \{i_1, i_2 ... i_m\}$ be a set of litterals called *items*.
- A sequence S is an ordered list of itemsets
- Sequence inclusions

Exemple

- $\mathcal{I} = \{a, b, c, d\}$
- $it_1 = (bcd)$, $it_2 = (ab)$
- $S =< (bcd)(ab) >$ (5-sequence)
- $< (bc)(a) > \preceq < (bcd)(ab) >$
- $< (a)(b) > \prec\succ < (b)(ab) >$

Definition (Support)

The support of a sequence S is defined as :

$$Support(S, \mathcal{D}) = \frac{|\{C \in \mathcal{D} | S \preceq C_{trans}\}|}{|\{C \in \mathcal{D}\}|}$$

Sequential Pattern mining

Extract all the frequent sequences S, i.e verifying :

$$Support(S, \mathcal{D}) \geq \sigma$$

with $0 \leq \sigma \leq 1$

Classical and incremental approaches

Classical approaches

1. Levelwise generate-and-prune :
   - SPADE : inverted database representation
   - SPAM : binary representation
2. Pattern-Growth :
   - PrefixSPAN : multiple database projection

Incremental approaches

Taking into account the dynamic evolution of a customer database ISE, ISM and IncSPAN (no deletion)

Remarks

1. Generation : Joint operations are known to be blocking operations [Babcock et al,2002]
2. There is more than 1 pass over $\mathcal{D}$ for all these algorithms however stream mining requires one-pass algorithms

Data streams approaches

- SMDS
- SPEED

Synopsis Construction

Requirements

- Broad Applicability
- One Pass Constraint
- Time and Space Efficiency
- Robustness
- Evolution sensitive

Techniques

1. Sampling Methods like Reservoir Sampling
2. Histograms
3. Wavelets
4. Sketches

Reservoir Sampling (Vitter 1985)

Main idea

An unbiased reservoir is maintained by probabilistic insertions and deletions

- Initialization : the first n points are directly added to the reservoir.
- When the $(t+1)^{th}$ point from the reservoir is received, it is added with a probability $\frac{n}{t+1}$ and replaces a random point in the reservoir.

Observations

- Insertion probabilities reduces with stream progression
- Unbiased reservoir maintained

Disadvantages

- The reservoir may not represent data stream evolutions
- Applications focusing on recent events from the data streams may get inaccurate results
- Smaller and smaller portions of the sample remains relevant with time

Biased Reservoir Sampling (Aggarwal 2006)

Main idea

- Use a temporal bias function to regulate the stream sample.
- This ensures that recent points from the data streams have higher probability to get inserted into the reservoir.

- Helps obtaining a biased and unbiased sample
- The bias is useful for applications focusing on representing the recent behaviour of the data streams

Observations

- An easy to use memory-less bias functions class is the *exponential bias functions* defined as :

$$f(r,t) = e^{-\lambda(t-r)}$$

  The parameter $\lambda \in [0,1]$ defines the bias rate
- The bias function is proportional to $p(r,t)$
- $p(r,t)$ is the probability that a point inserted at the instant r is still belonging to the reservoir when a point arrives at instant t
- In the special case of *exponential bias functions* the maximum reservoir requirement is bounded by $\frac{1}{\lambda}$ for small λ values

Challenges

- All classical mining algorithms have a strong hypothesis stating that a database can be loaded into main memory.
- What about real-world databases containing gigabytes of transactions ?
- Nowadays we can afford approximate solutions but can we assure bounds on the size of the samples given a desired accuracy ?

Sample size

- Error :

$$e(s, \mathcal{S_D}) = |Support(s, \mathcal{S_D}) - Support(s, \mathcal{D})|$$

X_i a random variable for the i^{th} customer with :

- $Pr[X_i = 1] = p_i$ if i^{th} customer supports the sequence s
- $Pr[X_i = 0] = 1 - p_i$, if not.

Note

We are in presence of Poisson trials as the number t of trials in which the probability of success p_i varies from trial to trial.

Sample size

- The number of customers in the sample that supports the sequence s :

$$X(s, \mathcal{S_D}) = \sum_i X_i = Support(s, \mathcal{S_D}) \times |\mathcal{S_D}|$$

- The expected number of customers that support the sequence s in the sample is :

$$E[X(s, \mathcal{S_D})] = Support(s, \mathcal{D}) \times |\mathcal{S_D}|$$

Theorem

Given a sequence s then $Pr[e(s, \mathcal{S_D}) > \epsilon] \leq \delta$ iff the reservoir size is :

$$|\mathcal{S_D}| \geq \ln(\frac{2}{\delta})\frac{1}{2\epsilon^2}$$

Proof sketch

1. Start from $Pr[|Support(s, \mathcal{S}_{\mathcal{D}}) - Support(s, \mathcal{D})| > \epsilon]$
2. introduce $X(s, \mathcal{S}_{\mathcal{D}})$ and $E[X(s, \mathcal{S}_{\mathcal{D}})]$
3. Use chernoff bounds to get the previous result

Observations

- We easily get an (ϵ, δ)-approximation
- Chernoff bound is not always very tight, but in this case it is acceptable
- We get samples of reasonable size with tolerable error :

ϵ	δ	$\mathcal{S}_{\mathcal{D}}$
0.01	0.01	26492
0.01	0.001	38005
0.001	0.01	2649160

Extending to Data Streams : the challenges

- We would like to approximate sequences support by maintaining a dynamic sample
- We would like to have both biased and unbiased sample (user-defined granularity)
- Use biased reservoir approach but with respect to our (ϵ, δ)-approximation

Analysis

We are working on biased reservoir samples, the following corollary gives an upper bound on the bias rate :

Corollary

Given an error bound ϵ and a maximum probability δ that $e(s, \mathcal{S}_\mathcal{D}) > \epsilon$ we get an upper bound on the bias rate :

$$\lambda \leq \frac{2\epsilon^2}{ln(2/\delta)}$$

- Proof sketch
  - $|\mathcal{S}_\mathcal{D}| \leq \frac{1}{1-e^{-\lambda}}$
  - $|\mathcal{S}_\mathcal{D}| \leq \frac{1}{\lambda}$
  - replace in the theorem

Observations

- The bias rate depends of the accuracy we want
- the accuracy of our mining results is optimal when the reservoir is full
- The reservoir maintained is very small in term of space requirements

ϵ	δ	λ	$\mathcal{S}_{\mathcal{D}}$
0.01	0.01	0.0000377	26492
0.01	0.001	0.0000263127	38005
0.001	0.0001	0.0000201949	49518

Algorithm

1. Check if customer C_i is present in the reservoir
2. If no, throw a coin
   - if Success ($< \frac{q}{n}$) add the customer to the reservoir
   - Else replace with a random position in the reservoir
3. If present in the reservoir then add C_i itemset

Plan
Motivations
Preliminary Concepts
Related Work
Sequential Pattern Mining
Synopsis Construction
Sampling in static databases
Extending to Data Streams
Experimental Results
Conclusion and Summary

Observations

We have to show that the replacement policy in the algorithm respects the exponential bias behaviour with $\lambda = \frac{1}{n}$

- Proof sketch
  - Probability that a customer is in the reservoir $\frac{1}{q}$
  - Probability to throw a customer is
$$(1-\frac{1}{q})(\frac{q}{n})(\frac{1}{q}) = \frac{q-1}{qn}$$
  - If the customer is inserted at the time r and is still in the reservoir at time t, then it did not get ejected in $t-r$ iterations : $(1-\frac{q-1}{qn})^{t-r}$
  - $(1-\frac{q-1}{qn})^{t-r} = [(1-\frac{q-1}{qn})^n]^{\frac{t-r}{n}}$
  - For large value of n, $(1-\frac{q-1}{qn})^n$ is approximately equal to $\frac{1}{e}$

Plan
Motivations
Preliminary Concepts
Related Work
Sequential Pattern Mining
Synopsis Construction
Sampling in static databases
Extending to Data Streams
Experimental Results
Conclusion and Summary

Experiments

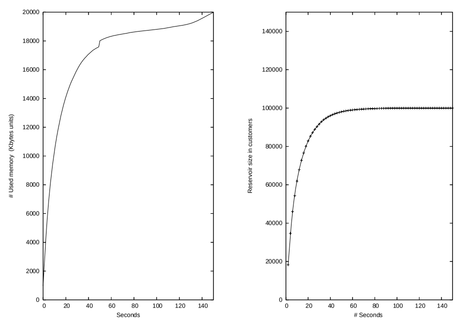

Fig.: Memory requirements for $\lambda = 0.00001$

Summary

- No sampling techniques for sequential patterns mining
- We introduced approximate approaches that work for mining on static databases and we extended it to data streams
- We get a biased sample, quality does not degrade with stream progression
- Extremely easy to implement and easy to maintain (small space requirements depending on bias rate defined by the user)

Thank you for your attention