Mining Maximal Sparse Interval

Naba Jyoti Sarmah
Dept. of Computer Science
Gauhati University
Guwahati-14

Anjana Kakoti Mahanta
Dept. of Computer Science
Gauhati University
Guwahati-14

ABSTRACT
Intervals are found in many real life applications such as web uses; stock market information; patient disease records; records maintained for occurrences of events, either man made or natural etc. Mining frequent intervals from such data allow us to group the transactions with similar behavior. Similar to frequent intervals, mining sparse intervals are also important. In this paper we define the notion of sparse and maximal sparse interval and also propose an algorithm for mining maximal sparse intervals. Computer programs were written and experimented on real life data set and results obtained have been reported. The correctness of the algorithm has also been proved.

General Terms
Interval Data Mining, Algorithms.

Keywords
Data Mining, Interval Data Mining, Maximal Sparse Interval.

1. INTRODUCTION
Data mining has received considerable amount of attention in last few years, many techniques have been developed for extracting information from huge amount of data generated by different domains. The most notable works on this field are association rule mining [1], sequential pattern mining [2] etc. These works are mainly carried out for market-basket data environment where each transaction is associated with a set of items. Here frequency of occurrences of items is used, to decide whether the pattern (which may be a relation among the items) is frequent or not. Many real world data are associated with a timestamp describing the occurrence time of the transaction. Many methods have been developed for mining temporal pattern [3], periodic pattern [4] from such data.

Many real world data are associated with durations instead of timestamps. A record in such dataset consists of a transaction and the starting and ending point of an interval in which the transaction had taken place. The transactions may be events and then the corresponding intervals may be the intervals in which the events had taken place. The duration of an event can be time, distance etc. In [5], Allen had first proposed Logic for representing interval data and a method for mining patterns from such datasets. In [6] the notion of maximal frequent interval is introduced and a method is proposed for mining maximal frequent intervals from such data using the concept of I-tree. Mining infrequent intervals play critical role in many situations. In [7] the authors had proposed a method for mining minimal infrequent intervals from multidimensional data.

In this paper we define maximal sparse intervals and propose a method for mining maximal sparse intervals. A sparse interval is an infrequent interval which does not properly contain any frequent interval except the empty interval which is always considered as frequent. A maximal sparse interval is a sparse interval which is not contained in any sparse interval. We feel that mining sparse intervals is also important for the reason discussed below-

Suppose a cellular phone company maintains records of time and length of all phone calls. Mining sparse intervals from such data will discover those intervals which are not frequent and by using these intervals company can either prepare some plan or some scheme to attract customers to these intervals or company can take decision to use the channel for some other purpose. The same is the case for web based learning system recording times at which each student logs on and off the system. Mining maximal sparse intervals enable the system administrator to discover the intervals during which very few numbers of students are online. If a student has the information that the time period $[t_1, t_2]$ is sparse, then he/she can plan to use the system for downloading or uploading any time period $[\overline{t1}, \overline{t2}]$ where $t_1 \leq \overline{t1} \leq \overline{t2} \leq t_2$. If one can extract the maximal sparse intervals then the corresponding decision makers can think of better utilization of these time periods.

In our work we have developed an efficient method for mining maximal sparse intervals after extracting the maximal frequent intervals and have tested it with some real life datasets.

In this paper in section 2 we discuss some recent work done in the field of interval data mining. In section 3 we formally define the problem and give some preliminary definitions. In section 4 we prove some properties of maximal sparse interval. In section 5 we discuss the proposed method for mining maximal sparse intervals. Result and discussion are given in section 6. Section 7 contains conclusion and lines for future work.

2. RELATED WORKS
Mining patterns from interval data is an active research area now. Many significant works has been carried out in this field. In [5] Allen proposed a temporal logic for maintaining the interval data where he defines thirteen possible relationships between two intervals. He also had proposed a method for mining temporal pattern from interval data. In [6] Lin proposed a method for mining maximal frequent interval using I-tree. He proposes a preorder traversal algorithm for mining maximal frequent interval from I-tree in $O(n^2)$ time. In [8] authors proposed a faster way for construction of I-tree. In [9] Kam & Fu further used the Allen's temporal logic and had defined a new method for finding the temporal patterns. In [10] authors proposed a new representation for interval based events and also had formulated a method for finding temporal patterns from interval based event data. In [7] a method was proposed for mining minimal infrequent intervals from multidimensional data. In [11] the authors have proposed a method for mining minimal infrequent intervals; the method

uses the maximal frequent intervals of the dataset to mine minimal infrequent intervals.

3. PRELIMINARIES

Let D be an interval dataset with n transactions, where each transaction t_i ($1 \le i \le n$), has associated with it an interval $[\ell,r]$ over a discrete domain, where ℓ is the starting point of the transaction and r is the end point of the transaction. Each transaction may be an event that has occurred in the time interval. Our problem is to mine all the maximal sparse intervals within our domain of interest. Let ℓ_{min} be the smallest left endpoint and r_{max} be the largest right endpoint in this domain. Since our domain is a discrete domain; for any point ℓ in the domain where $\ell \neq r_{max}$, $\ell+1$ will denote the immediate next point of ℓ in the domain. Similarly ℓ -1 where $\ell \neq l_{min,}$will denote the immediate previous point of ℓ in the domain. For a given transaction t and an interval [a,b], we have the following definitions.

Transaction: A transaction in an interval dataset contains an interval $[\ell,r]$ over a discrete domain along with the other attributes of the transaction. Here ℓ is the starting point of the transaction and r is the ending point of that transaction.

Support of an Interval: A transaction 't' supports an interval [a,b] if $[a,b] \subseteq [\ell_t , r_t]$ ie, if $\ell_t \le a \le b \le r_t$, where ℓ_t is the starting time and r_t is the ending time of the interval associated with t. For a given interval [a,b], sup([a,b]) will denote the number of transactions in D that supports [a,b].

Frequent Interval: For a given minimum support threshold min_sup, with 0<min_sup<n, an interval is called frequent if its support is greater than or equal to min_sup. Obviously if $[\ell,r]$ is frequent, then $\ell_{min} \le \ell \le r \le r_{max}$.

Infrequent Interval: An interval $[\ell,r]$ will be called infrequent if $\ell_{min} \le \ell \le r \le r_{max}$ and it is not frequent.

Sparse Interval: A sparse interval is an infrequent interval which does not properly contain any frequent interval except the empty interval which is always considered as frequent, i.e. if an interval $[\ell,r]$ is a sparse interval and $[\ell', r'] \subset [l,r]$ where $\ell' \le r'$, then $[\ell', r']$ is infrequent.

Maximal Sparse interval: A maximal sparse interval is a sparse interval which is not contained in any sparse interval, i.e. if an interval $[\ell,r]$ is a maximal sparse interval and $[\ell,r] \subset [\ell', r']$ then $[\ell', r']$ is not a sparse interval.

Example:

Table 1: Dataset for example

Tid	1	2	3	4	5	6	7	8	9
Intervals	[1,5]	[1,4]	[2,6]	[2,5]	[6,10]	[6,11]	[7,12]	[8,12]	[8,13]

Consider the dataset shown in the table 1, suppose all the endpoints of the intervals are in discreate domain
D = { v | 1 ≤ v ≤ 13 and v is an integer }
If we consider the minimum support threshhold as 4 we have the following results-
Maximal Frequent Intervals : [2,4], [7,10], [8,11];
Maximal Sparse Intervals : [1,1], [5,6], [12,13];

4. SOME PROPERTIES OF THE MAXIMAL FREQUENT INTERVAL AND MAXIMAL SPARSE INTERVAL

In this section we prove certain properties of maximal sparse intervals. Let ℓ_{min} and r_{max} be as defined in section 3. If $[\ell,r]$ and $[\ell',r']$ are two distinct maximal frequent intervals then we have either (i) $\ell < \ell'$ and $r < r'$ or (ii) $\ell' < \ell$ and $r' < r$; since otherwise one interval will contain the other interval. Hence the maximal frequent intervals $[\ell_1, r_1]$, $[\ell_2, r_2]$, … ,$[\ell_k, r_k]$ in D can be arranged as $\ell_1 < \ell_2 < \ldots < \ell_k$ and $r_1 < r_2 < \ldots < r_k$, where k is assumed to be the number of such intervals in D. Because of this for any 'a' with $\ell_{min} \le a \le r_{max}$ there can be at most one maximal frequent interval with 'a' as a left end point. We have the following theorems for maximal sparse intervals.

Theorem 1: Every sparse interval is contained in some maximal sparse interval

Proof: If $[\ell, r]$ be a sparse interval. If $[\ell, r]$ is not a maximal sparse interval then $[\ell,r] \subset [\ell',r']$, where $[\ell',r']$ is a sparse interval. If $[\ell', r']$ is not a maximal sparse interval then $[\ell', r'] \subset [\ell'', r'']$, where $[\ell'', r'']$ is a sparse interval. Since only a finite number of intervals can be there with endpoints between ℓ_{min} and r_{max}, the above process cannot continue infinitely and there will be some maximal sparse interval $[\ \bar{l}, \bar{r}\ ]$ such that $[\ell,r] \subseteq [\ \bar{l}, \bar{r}\ ]$. Hence every sparse interval is contained in some maximal sparse interval.

Theorem 2: if $\ell_{min} < \ell_1$ then $[\ell_{min}, \ell_1 -1]$ is a maximal sparse interval.

Proof: From the definition of sparse interval, $[\ell_{min}, \ell_1-1]$ will be a sparse interval if
(i) $[\ell_{min}, \ell_1 -1]$ is an infrequent interval and
(ii) All non-empty subintervals of $[\ell_{min}, \ell_1-1]$ are infrequent.
Since every frequent interval is contained in a maximal frequent interval and ℓ_1 is the lowest left endpoint of all the maximal frequent intervals of the dataset, there is no frequent interval starting at some ℓ' where $\ell' < \ell_1$ and so $[\ell_{min}, \ell_1-1]$ is an infrequent interval. Also none of the subintervals of $[\ell_{min}, \ell_1-1]$ is frequent for the same reason. Hence $[\ell_{min}, \ell_1-1]$ is a sparse interval.

Since ℓ_{min} is the lowest left end point in our domain, any interval properly containing $[\ell_{min}, \ell_1-1]$ will have to be of the form $[\ell_{min}, \ell']$ for some $\ell' \ge \ell_1$. But $[\ell_{min}, \ell']$ cannot be sparse since $[\ell_1, \ell_1]$ is frequent (since it is contained in $[\ell_1, r_1]$). Therefore $[\ell_{min}, \ell_1-1]$ is a maximal sparse interval.

Theorem 3: If $r_{max} > r_k$ then $[r_k+1, r_{max}]$ is a maximal sparse interval.

Proof: From the definition of sparse interval, $[r_k, r_{max}]$ will be a sparse interval if
(i) $[r_k+1, r_{max}]$ is an infrequent interval and
(ii) All non-empty subintervals of $[r_k+1, r_{max}]$ are infrequent.
Since every frequent interval is contained in a maximal frequent interval and r_k is the largest right endpoint of all the maximal frequent intervals of the dataset, there is no frequent interval ending at some r' where $r' > r_k$ and so $[r_k+1, r_{max}]$ is an infrequent interval. Also none of the subintervals of $[r_k+1, r_{max}]$ is frequent for the same reason. Hence $[r_k+1, r_{max}]$ is a sparse interval.

Since r_{max} is the largest right end point in our domain, any interval properly containing $[r_k+1, r_{max}]$ will have to be of the form $[r',r_{max}]$ for some $r' \leq r_k$. But $[r',r_{max}]$ cannot be sparse since $[r_k,r_k]$ is frequent (since it is contained in $[l_k,r_k]$). Therefore $[r_k+1, r_{max}]$ is a maximal sparse interval.

Theorem 4: If $r_i + 1 \leq \ell_{i+1} - 1$ then $[r_i+1, \ell_{i+1}-1]$ is a maximal sparse interval.

Proof: Since $r_i + 1 \leq \ell_{i+1} - 1$ the interval $[r_i+1, \ell_{i+1}-1]$ is non-empty. From definition $[r_i+1, \ell_{i+1}-1]$ will be a sparse interval if

(i) $[r_i+1, \ell_{i+1}-1]$ is an infrequent interval and
(ii) All non-empty subintervals of $[r_i+1, \ell_{i+1}-1]$ are infrequent.

Since every frequent interval is contained in a maximal frequent interval and $[r_i+1, \ell_{i+1}-1]$ is not contained in any $[\ell_j, r_j]$ for all $1 \leq j \leq i$ and also in any $[\ell_{j'}, r_{j'}]$ for all $i+1 \leq j' \leq k$, it is an infrequent interval. Also all its subintervals are infrequent for the same reason. Hence $[r_i+1, \ell_{i+1}-1]$ is a sparse interval.

Again let $[\ell', r']$ be any interval properly containing $[r_i+1, \ell_{i+1}-1]$. Then $[\ell', r']$ will contain the point r_i or the point ℓ_{i+1} or both. In the first case $[r_i, r_i]$ is a frequent interval contained in $[\ell', r']$ and in the second case $[\ell_{i+1}, \ell_{i+1}]$ is a frequent interval contained in $[\ell', r']$ and in the third case both the frequent intervals $[r_i, r_i]$ and $[\ell_{i+1}, \ell_{i+1}]$ are contained in $[\ell', r']$. Hence $[\ell', r']$ cannot be a sparse interval.
Hence $[r_i+1, \ell_{i+1}-1]$ is a maximal sparse interval.

Theorem 5 proves the completeness of our proposed algorithm for mining maximal sparse intervals. For the proof of the theorem we need the following two lemmas.

Lemma 1. For any ℓ where $\ell_{min} \leq \ell \leq r_{max}$, ℓ cannot belong to both a sparse interval and a frequent interval.

Proof: This is because if ℓ is in a frequent interval then $[\ell, \ell]$ is a frequent interval and this implies that ℓ cannot be in a sparse interval as all subintervals of a sparse interval are infrequent.

Lemma 2. The maximal sparse intervals are mutually disjoint.
Proof: If possible let $[\ell', r']$ and $[\ell'', r'']$ be any two maximal sparse intervals having non-empty intersection. Since both the intervals are maximal, none is contained within the other. Without loss of generality we can assume that $\ell'' \leq r'$. Now let us consider the interval $[\ell', r'']$. This interval is infrequent since $[\ell', r']$ is an infrequent subinterval of it. Also all its subintervals are infrequent since any such sub interval is contained in $[\ell', r']$ or in $[\ell'', r'']$ or in both. Therefore $[\ell', r'']$ is a sparse interval. This contradicts our assumption that $[\ell', r']$ and $[\ell'', r'']$ are maximal sparse intervals.

Theorem 5: The maximal sparse intervals given by theorem 2, 3 and 4 are the only maximal sparse intervals in D.

Proof: Theorem 2 gives us a maximal sparse interval containing all points ℓ where $\ell < \ell_1$. Since any point can be at most in one maximal sparse interval [lemma 2], there cannot be any more maximal sparse interval in the interval $[\ell_{min}, \ell_1-1]$. Similarly theorem 3 gives us a maximal sparse interval containing all points ℓ where $\ell > r_k$. Since any point can be at most in one maximal sparse interval [lemma 2], there cannot be any more maximal sparse interval in the interval $[r_k+1, r_{max}]$.

For any two consecutive maximal frequent intervals $[\ell_i, r_i]$ and $[\ell_{i+1}, r_{i+1}]$ in the sorted order as discussed earlier in this section if we have $r_i+1 < \ell_{i+1}-1$ then theorem 4 gives us a maximal sparse interval containing all points ℓ such that $r_i < \ell < \ell_{i+1}$. Lemma 2 states that a point cannot belong to two maximal sparse intervals. So using lemma 1 and lemma 2 it is easy to see that there cannot be any more maximal sparse interval in the domain D which actually is the interval $[\ell_{min}, r_{max}]$.

5. ALGORITHM PROPOSED AND COMPLEXITY ANALYSIS

The following algorithm is proposed to extract maximal sparse intervals in O(n) time if the maximal frequent intervals of the dataset are provided.

Algorithm: Mining Maximal Sparse Intervals
Input: Maximal frequent intervals in sorted order of their endpoints
Output: Maximal Sparse Intervals

```
Step1.  MSI = empty
Step2.    Read the first maximal frequent interval [ℓ1,r1]
Step3.    if (ℓmin < ℓ1 )
Step4.      append [ℓmin, ℓ1 - 1] to MSI
Step5.  lc = r1
Step6.  Repeat till the end of list of maximal frequent
        intervals is reached
Step7.      Read the next maximal frequent interval < ℓ,r>
Step8.      if ( lc+1 ≤ ℓ-1) )
Step9.            append [lc+1, ℓ-1] to MSI
Step10.     endif
Step11.   lc=r
Step12. end repeat
Step13. if ( lc  < rmax )
Step14.   append [lc+1, rmax] to MSI
Step15. endif
```

Complexity Analysis of the algorithm

After obtaining the maximal frequent intervals in $O(n^2)$ time as proposed in [6], the maximal sparse intervals are extracted by moving through the list of maximal frequent intervals once, using ℓ_{min} and $r_{max.}$ Another scan through the database is not required. Extraction of one maximal sparse interval requires O(1) time. Both ℓ_{min} and r_{max} can be obtained in linear time. The number of maximal frequent intervals is linear in the size of input data since each such interval has a distinct left endpoint and these points are from the left endpoints in the input data. Also, since the number of maximal sparse intervals is more than that of the maximal frequent intervals at most by a constant value, their number is also linear in size of the input data. Therefore after obtaining the maximal frequent intervals the maximal sparse intervals can be obtained in O(n) time.

6. RESULT AND DISCUSSION

To test the proposed algorithm we have developed programs in C++ and have used the dataset obtained from "Bodhidroom" (www.bodhidroom.idolgu.org); the online e-learning portal of Institute of Distance and Open Learning, Gauhati University. The records contain the log file of the website containing information about the login and logout time of the users. The dataset contains a total of 10031 records from 31/3/2009 to 14/10/2011. Mining maximal sparse interval enables the system to find out the intervals during which less than a predefined number of users are online. This

information can be used for various purposes to enhance the performance of the system.

The results are found as follows.

Table 2: Experimental Result

Threshold	No of maximal frequent interval	No of maximal sparse interval	Time in second
1	8070	7180	0.000713
2	2358	2052	0.000199
3	554	493	0.000080
4	108	90	0.000012
5	30	24	0.000006
6	7	4	0.000004

Sample input file: data_sec.txt

login time	**logout time**
31-3-2009 21:14:47	31-3-2009 21:45:27
31-3-2009 21:45:47	31-3-2009 22:48:56
31-3-2009 22:49:58	31-3-2009 22:52:56
31-3-2009 22:53:11	31-3-2009 22:59:37
31-3-2009 23:0:36	31 3-2009 23:9:54
31-3-2009 23:10:13	31 3-2009 23:32:18
1-4-2009 10:18:3	1-4-2009 10:29:1
8-4-2009 19:43:0	8-4-2009 20:0:43
8-4-2009 20:1:2	8-4-2009 20:2:41
8-4-2009 20:6:0	8-4-2009 20:15:22
8-4-2009 20:17:33	8-4-2009 20:28:56
8-4-2009 20:30:59	9-4-2009 0:28:25
……..	……

Threshold :: 4

Maximal Frequent Intervals obtained from the dataset::

left_end points	**right_end points**
5-11-2009 16:30:2	5-11-2009 16:31:43
11-11-2009 15:59:48	11-11-2009 16:9:9
14-11-2009 12:59:14	14-11-2009 13:1:22
16-11-2009 13:32:15	16-11-2009 13:33:9
16-11-2009 13:33:18	16-11-2009 13:57:6
16-11-2009 13:43:19	16-11-2009 14:1:40
16-11-2009 14:2:11	16-11-2009 14:11:46
16-11-2009 14:12:25	16-11-2009 14:12:55
16-11-2009 14:13:4	16-11-2009 14:24:17
16-11-2009 14:32:41	16-11-2009 14:44:20
16-11-2009 14:45:17	16-11-2009 14:46:37
19-11-2009 16:7:5	19-11-2009 16:14:13
19-11-2009 16:25:7	19-11-2009 16:26:38
19-11-2009 16:29:58	19-11-2009 16:30:49
…….	…….

Maximal Sparse Intervals detected by the algorithm::

left_end points	**right_end points**
31-3-2009 21:14:47	5-11-2009 16:30:1
5-11-2009 16:31:44	11-11-2009 15:59:47
11-11-2009 16:9:10	14-11-2009 12:59:13
14-11-2009 13:1:23	16-11-2009 13:32:14
16-11-2009 13:33:10	16-11-2009 13:33:17
16-11-2009 14:1:41	16-11-2009 14:2:10
16-11-2009 14:11:47	16-11-2009 14:12:24
16-11-2009 14:12:56	16-11-2009 14:13:3
16-11-2009 14:24:18	16-11-2009 14:32:40
16-11-2009 14:44:21	16-11-2009 14:45:16
16-11-2009 14:46:38	19-11-2009 16:7:4
19-11-2009 16:14:14	19-11-2009 16:25:6
19-11-2009 16:26:39	19-11-2009 16:29:57
…….	…….

7. CONCLUSION AND LINES FOR FUTURE WORK

The notion of maximal sparse interval has been introduced here and an algorithm has been proposed for mining maximal sparse intervals from an interval based dataset for a given minimum support threshold. The proposed algorithm generates a set of maximal sparse intervals using the maximal frequent intervals in the dataset. The effectiveness of the proposed algorithm was tested for real life dataset obtained from IDOL. The proposed method depends on a maximal frequent interval mining algorithm to mine maximal sparse intervals and the method for mining maximal frequent interval proposed in [6] is an $O(n^2)$ algorithm.

Possibility of finding the maximal sparse intervals directly without finding the maximal frequent intervals may be looked into. More work could be done to extend the work to multi dimensional space. Further works can be carried out to improve the performance of the system. If the end points of the intervals are calendar dates then the possibility of finding periodicity of patterns is another line of future work to be studied.

8. ACKNOWLEDGMENTS

The first author is an INSPIRE fellow, fellowship is granted by Department of Science and Technology for pursuing his Ph D.

9. REFERENCES

[1] Rakesh Agrawal, Ramakrishnan Srikant, "Fast Algorithm for Mining Association Rules", Proceedings of the 20th VLDB conference, Santiago, Chile, 1994.

[2] Rakesh Agrawal, Ramakrishnan Srikant, "Mining Sequential Patterns", Proceedings of the Eleventh International Conference on Data Engineering, p.3-14 March 06-10, 1995.

[3] A. K. Mahanta, N. H. Son, "Mining Interesting Periodicities of Temporal Patterns" Proceedings of IPMU'08, p.1757-1764, June 22-27, 2008.

[4] A. K. Mahanta, F. A. Mazarbhuya, H. K. Baruah, "Finding calendar-based periodic patterns" Pattern Recognition Letters, p.1274-1284, Vol 29 Issue 9, July 2008.

[5] J. F. Alen, "Maintaining Knowledge about Temporal Intervals" Communications of the ACM, Vol 26, Nov 1983.

[6] Jun-Lin Lin, "Mining Maximal Frequent Intervals", Proceedings of 2003 ACM symposium on Applied Computing, p.426-431, ACM, New York(2003).

[7] Khaled M. Elbassioni, "Finding All Minimal Infrequent Multi-dimensional Intervals", Proceedings of the 7th Latin American conference on Theoretical Informatics, p. 423-434, 2006.

[8] M. Dutta, A. K. Mahanta, " An Efficient Method for Construction of I-tree", Proceedings of National Workshop on Design and Analysis of Algorithm(NWDA)2010.

[9] Po-shan Kam, Ada Wai-chee Fu, "Discovering Temporal Patterns for Interval-based Events" Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery, p.317-326, 2000.

[10] Shin-Yi Wu, Yen-Liang Chen, "Mining Nonambiguous Temporal Patterns for Interval-Based Events" IEEE Transactions on Knowledge and Data Engineering, Vol 19 No 6, June 2007.

[11] D. I. Mazumdar, D. K. Bhattacharyya, M. Dutta, "Mining Minimal Infrequent Intervals", Journal of Computer Science and Engineering, communicated.