Consecutive Retrieval Property - Revisited

Jitender S. Deogun*
Department of Computer Science and Engineering
University of Nebraska – Lincoln
Lincoln, NE 68588

K. Gopalakrishnan
Department of Mathematics
East Carolina University
Greenville, NC 27858

Abstract: *The connection between the consecutive retrieval property and interval graphs is explored. Necessary and sufficient conditions for the existence of the consecutive retrieval property are developed from the standpoint of graph theory. These conditions are based on a characterization of unit interval graphs developed in this paper.*

Keywords: *Consecutive Retrieval Property, Unit Interval Graphs, Information Retrieval.*

1 Introduction

In 1972, Ghosh [6, 7] introduced the concept of the consecutive retrieval file organization. Easwaran [3, 4] developed a graph theoretic approach for analyzing the consecutive retrieval property and established some necessary conditions for its existence. Our main objectives are to clarify some of the myths regarding the relationship between the consecutive retrieval property and interval graphs and to develop necessary and sufficient conditions for the existence of the consecutive retrieval property from the standpoint of graph theory.

A storage media in which the records are searched in a linear pattern is called a linear storage. Efficient organization of files in linear storage medium is our concern in this paper. Suppose $R = \{r_1, r_2, \dots, r_m\}$ is a set of records and $Q = \{q_1, q_2, \dots, q_n\}$ is a set of queries. It is assumed that it is possible to determine unambiguously whether or not a given record is relevant to a given query.

A query set Q is said to have the *Consecutive Retrieval Property* (CRP) with respect to a record set R if there exists an organization of R without redundancy such that for every query in Q, the pertinent records are stored in consecutive storage locations. Such an organization is called *Consecutive Retrieval File Organization* (CRFO).

$$M = \begin{array}{cc} & \begin{array}{ccccc} r_1 & r_2 & r_3 & \dots & r_m \end{array} \\ \begin{array}{c} q_1 \\ q_2 \\ \vdots \\ q_n \end{array} & \begin{vmatrix} 1 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 1 \end{vmatrix} \end{array}$$

Figure 1: The incidence matrix of a query set.

Under the assumptions, it is possible to associate an incidence matrix with Q and R. Let M be an $n \times m$ matrix whose n rows correspond to the n queries of Q and whose m columns correspond to the m records of R. $M(i, j)$ contains 1 if r_j is pertinent to q_i and 0 otherwise (see Fig. 1).

The CRP between Q and R can be stated in terms of properties of M [6]. The CRP exists between Q and R iff there exists a permutation of the columns of M for which the 1's in every row of M are in consecutive positions.

We can assume without loss of generality that each query in Q has at least two pertinent records. This is because singleton queries do not affect CRP [4]. Further we assume that Q is a set of non-compromising subsets that cover R, that is, if $i \neq j$

*Supported in part by the Army Research Office, Grant No. DAAH04-96-1-0325, under DEPSCoR program of Advanced Research Projects Agency, Department of Defence.

then neither q_i is a subset of q_j nor q_j is a subset of q_i. Non-compromising subsets arise naturally in classification schemes and level clustering [2].

The rest of the paper is organized as follows. In Section 2, an earlier connection between interval graphs and the CRP is shown to be incorrect. In Section 3, we introduce the notion of equivalent graphs and develop necessary conditions for the existence of the CRP. In Section 4, a characterization of unit interval graphs is presented. In Section 5, this characterization is used to develop the main result viz., a graph theoretical necessary and sufficient condition for the existence of the CRP. In Section 6, we present the concluding remarks and discuss some open problems.

2 CRP and Interval Graphs

We investigate an earlier connection between the CRP and interval graphs. A counterexample to show that the earlier connection is incorrect is presented. Moreover, we show that no such connection could exist.

The *intersection graph* of a set of queries Q, denoted by $\Gamma(Q)$, is defined as follows: for each query q_i in Q, there exists a corresponding vertex v_i in $\Gamma(Q)$ and vice versa and for $i \neq j$, v_i is joined with v_j by an edge iff $q_i \cap q_j \neq \phi$.

The following theorem is stated in *Data Base Organization for Data Management* [8, Theorem 6.3.2, p. 247].

Theorem 2.1. *[8] A necessary and sufficient condition for the existence of the CRP of a query set Q with respect to a record set R is that its intersection graph be an interval graph.*

However, the theorem does not appear in the second edition of the book; but no elaboration was provided. We show that the theorem is incorrect and develop a correct necessary and sufficient condition for the existence of the CRP.

Suppose $R = \{r_1, r_2, r_3\}$ and $Q = \{q_1, q_2, q_3\}$. Let $q_1 = \{r_1, r_2\}$, $q_2 = \{r_2, r_3\}$ and $q_3 = \{r_1, r_3\}$. Then intersection graph $\Gamma(Q)$ is the triangle shown in Fig. 2.

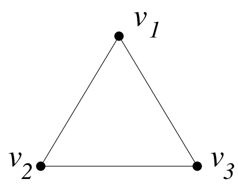

Figure 2: The intersection graph of the counterexample

The above graph, as is any complete graph, is an interval graph. However, it can be easily verified that the above set Q of queries does not have the CRP. This is clearly a counterexample to the above theorem. The condition stated in the theorem is only a necessary condition and not a sufficient one.

Moreover, there is a loss of information when we move from the query-record pattern to its intersection graph. As a result, it is impossible to determine whether a query-record pattern has the CRP just based on its intersection graph. We formalize this idea in the form of the following theorem.

Theorem 2.2. *Any necessary and sufficient condition for the existence of the CRP of a query set Q with respect to a record set R that is based only on its intersection graph is false.*

Proof. To prove this, we exhibit two query-record patterns such that both have the same intersection graph but one of them has the CRP and the other one does not.

Let $Q_1 = \{q_1, q_2, q_3\}$ and $R_1 = \{r_1, r_2, r_3\}$. Let $q_1 = \{r_1, r_2\}$, $q_2 = \{r_2, r_3\}$ and $q_3 = \{r_3, r_1\}$. It is easily verified that (Q_1, R_1) does not have the CRP.

Let $Q_2 = \{q_1, q_2, q_3\}$ and $R_2 = \{r_1, r_2, r_3, r_4, r_5\}$. Let $q_1 = \{r_1, r_2, r_3\}$, $q_2 = \{r_2, r_3, r_4\}$ and $q_3 = \{r_3, r_4, r_5\}$. Clearly, (Q_2, R_2) has the CRP and a CRFO is $(r_1, r_2, r_3, r_4, r_5)$.

Note further that intersection graphs $\Gamma(Q_1)$ and $\Gamma(Q_2)$ are isomorphic to the triangle shown in Fig. 2. So any characterization based only on intersection graphs would yield the erroneous result that either both (Q_1, R_1) and (Q_2, R_2) have the CRP or both (Q_1, R_1) and (Q_2, R_2) do not have the CRP. $\square$

3 A Necessary Condition for Existence of the CRP

We develop a necessary condition for the existence of the CRP. The *equivalent graph* of Q, denoted by $\psi(Q)$, is defined as follows: for each record r_i in R, there exists a corresponding vertex v_i in $\psi(Q)$ and vice versa and for $i \neq j$, v_i is adjacent to v_j iff r_i and r_j are both relevant to some query q_k belonging to Q. Fig. 3 shows an example of Q-R matrix and Fig. 4 shows the corresponding equivalent graph $\psi(Q)$.

$$M = \begin{array}{c} \\ q_1 \\ q_2 \\ q_3 \end{array} \begin{array}{c} r_1 \quad r_2 \quad r_3 \quad r_4 \quad r_5 \quad r_6 \\ \left| \begin{array}{cccccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{array} \right| \end{array}$$

Figure 3: An example of a Q-R matrix.

Theorem 3.1. *A necessary condition for the existence of the CRP of a query set Q with respect to a record set R is that incidence matrix M be identical to the maximal cliques vs. vertices matrix of the equivalent graph $\psi(Q)$.*

Proof. We prove the contrapositive i.e., if M is not identical to the maximal cliques vs. vertices matrix of the equivalent graph $\psi(Q)$ then the CRP does not exist.

By construction of $\psi(Q)$, it is clear that each row of M correspond to a clique of $\psi(Q)$, not necessarily maximal. Let M' be the maximal cliques vs. vertices matrix of the equivalent graph $\psi(Q)$. Suppose a query q_i doesn't correspond to a maximal clique. Let $r_{i'}$ and $r_{j'}$ be two records pertinent to q_i. Note that each query has at least two pertinent records. The maximal clique that contains the clique corresponding to query q_i obviously should contain at least one record not pertinent to q_i. Let that record be $r_{k'}$. Since $v_{k'}$ is joined to $v_{i'}$, both

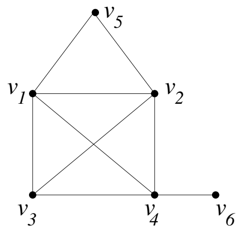

Figure 4: Equivalent graph $\psi(Q)$ of M in Fig. 3

$r_{i'}$ and $r_{k'}$ should appear in some query, say q_j. Similarly since $v_{k'}$ is joined to $v_{j'}$, both $r_{k'}$ and $r_{j'}$ should appear in some query say, q_k. Thus the matrix shown in Fig. 5 is a submatrix of M.

$$M = \begin{array}{c} \\ q_i \\ q_j \\ q_k \end{array} \begin{array}{c} r_{i'} \quad r_{j'} \quad r_{k'} \\ \left| \begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{array} \right| \end{array}$$

Figure 5: A submatrix of M

It can be verified that $\{q_i, q_j, q_k\}$ does not have the CRP. Consequently, it follows that Q does not have the CRP. Note that because of our assumption of non-compromising queries, we are able to select records $r_{i'}$ and $r_{j'}$ such that q_j and q_k are distinct queries. We have shown that every query is a maximal clique.

Now, suppose a maximal clique C of $\psi(Q)$ does not correspond to a query. By construction of $\psi(Q)$ it is clear that any edge of C must join two vertices whose corresponding records must be pertinent to some query in Q. C clearly must have more than two vertices. Since C does not correspond to query, it is not too difficult to see that there must exist three distinct records $r_{i'}, r_{j'}, r_{k'}$ that are not relevant to a common query. This would mean that the matrix shown in Fig. 5 would be a submatrix of M, so Q cannot have the CRP. Note again that our assumption of non-compromising queries ensures that $q_j \neq q_k$. Hence matrix M is identical to matrix M'. $\square$

4 A Characterization of Unit Interval Graphs

The intersection graph of a family of intervals of unit length on a real line is called a *Unit Interval Graph* (UIG). Several characterizations of unit interval graphs are known [9, 11]. We develop yet another characterization, which appears to be new.

It is well known that G is an unit interval graph iff G is an interval graph containing no induced copy of $K_{1,3}$ [9]. Further, an undirected graph G is an interval graph iff its maximal cliques vs. vertices matrix M has the consecutive 1's property for columns [5]. Using these facts, we develop a characterization of unit interval graphs.

Theorem 4.1. *Let $G = (V, E)$ be an undirected graph. Let M denote the maximal cliques vs. vertices matrix of G. Then G is a unit interval graph iff the rows and columns of M can be permuted in such a way that the 1's in each column and each row occur consecutively i.e., G is a unit interval graph iff M has the consecutive 1's property for both rows and columns.*

Proof. **Part 1.** First we will show that if M has the consecutive 1's property for both rows and columns, then G is an unit interval graph.

Since M has the consecutive 1's property for columns, it obviously follows that G is an interval graph. If G is not a unit interval graph, then it should contain an induced copy of the graph $K_{1,3}$ shown in Fig. 6.

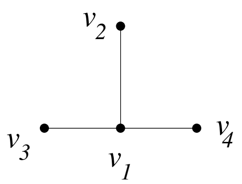

Figure 6: The graph $K_{1,3}$

Let C_2 be a maximal clique containing v_1 and v_2. Similarly, let C_3 be a maximal clique containing v_1 and v_3 and C_4 be a maximal clique containing v_1 and v_4. Cleary, C_2 contains neither v_3 nor v_4, C_3 contains neither v_2 nor v_4, and C_4 contains neither v_2 nor v_3. Hence matrix M' shown in Fig. 7 will be a submatrix of M.

$$M' = \begin{array}{c} \\ c_2 \\ c_3 \\ c_4 \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & v_3 & v_4 \end{array} \\ \left| \begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array} \right| \end{array}$$

Figure 7: A submatrix of M

It can be verified that no permutation of the columns of M' will ensure consecutive 1's in all the rows. Thus M' does not have the consecutive 1's property for rows. It follows that M does not have the consecutive 1's property for rows, and that is a contradiction. Hence our assumption that G is not a unit interval graph is wrong.

Thus if M has the consecutive 1's property for both rows and columns, then G is an unit interval graph.

Part 2. We show that if G is an unit interval graph then M has the consecutive 1's property for both rows and columns.

Since G is an interval graph, it follows that M has the consecutive 1's property for columns. It remains to show that M has the consecutive 1's property for rows.

Consider a unit interval diagram of G. Define $v_i \prec v_j$, if the left end of the interval corresponding to v_i is to the left of the left end of the interval corresponding to v_j or if the intervals corresponding to v_i and v_j are identical and $i < j$. Clearly, $\prec$ is a total ordering of the vertices. Now arrange the columns of M as per this total ordering of vertices. We claim that in this arrangement, 1's in each row occur consecutively.

If not, in some row a 1 is followed by a nonempty sequence of zeroes and then again by a 1, i.e., some row contains the sequence shown in Fig. 8.

Since v_{i_1} and v_{i_k} are in some clique, their intervals should intersect. Further, by construction, the interval corresponding to v_{i_2} starts (at the same point or) after the starting point of the interval cor-

$$\begin{array}{ccccc} v_{i_1} & v_{i_2} & v_{i_3} & \dots & v_{i_k} \\ 1 & 0 & 0 & \dots & 1 \end{array}$$

Figure 8: A subsequence of some row of M

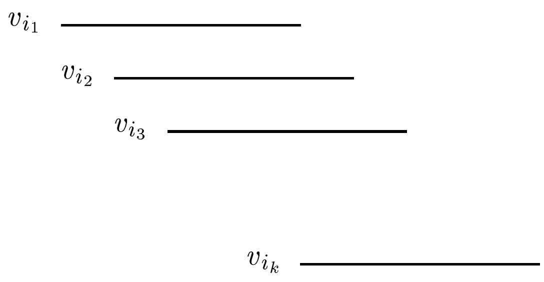

Figure 9: Unit Interval Diagram

responding to v_{i_1} but (at the same point or) before the starting point of the interval corresponding to v_{i_k}. Also the interval corresponding to v_{i_2} ends (at the same point or) after the ending point of the interval corresponding to v_{i_1} since all the intervals are of unit length. Hence the interval corresponding to v_{i_2} is a superset of the portion of the real line common to the intervals corresponding to v_{i_1} and v_{i_k} (see Fig. 9). It follows that no clique can contain v_{i_1} and v_{i_k} but not v_{i_2}. The same argument holds for v_{i_3},v_{i_4} $\dots$ $v_{i_{k-1}}$. Thus such a sequence cannot exist in any row, so M has the consecutive 1's property for rows. □

5 A Graph Theoretic Characterization of the CRP

We use the characterization of unit interval graphs to develop necessary and sufficient conditions for the existence of the CRP of a given Q-R matrix from the standpoint of graph theory.

Theorem 5.1. *A given Q-R matrix has the CRP iff it corresponds to the maximal cliques vs. vertices matrix of some unit interval graph. In other words, for the existence of the CRP for a Q-R matrix it is necessary and sufficient that the Q-R matrix be identical to the maximal cliques vs. vertices matrix of the equivalent graph $\psi(Q)$ and that $\psi(Q)$ is a unit interval graph.*

Proof. If the Q-R matrix corresponds to the maximal cliques vs. vertices matrix of some unit interval graph, then it follows from our previous characterization of unit interval graphs that the matrix has the consecutive 1's property for both rows and columns. Since the matrix has the consecutive 1's property for rows, it follows that the CRP exists between Q and R.

If the CRP exists between Q and R, then it follows from Theorem 3.1 that the Q-R matrix is identical to the maximal cliques vs. vertices matrix of the equivalent graph $\psi(Q)$. It also follows that the matrix has the consecutive 1's property for rows. It remains to be shown that the matrix has the consecutive 1's property for columns and therefore $\psi(Q)$ is indeed a unit interval graph.

Arrange the columns of M such that in each row 1's occur consecutively. Define $q_1 \prec q_2$ if the first 1 in the row corresponding to q_1 appears to the left of the first 1 in the row corresponding to q_2. In view of our assumption that no query is a superset or subset of another query, it is clear that $\prec$ is a total ordering of the queries. Now arrange the rows of M as per this total ordering of queries. We claim that in this arrangement 1's in each column occur consecutively.

If not, in some column a 1 is followed by a nonempty sequence of zeroes and then again by a 1. i.e., some column contains the sequence shown in Fig. 10.

Now all 1's in the row corresponding to q_{i_2} appear either before column j or after column j, as consecutive 1's in rows is already assured. Also, the first 1 in the row corresponding to q_{i_1} appears to the left of the first 1 in the row corresponding to q_{i_2}. If all the 1's in the row corresponding to q_{i_2} appear before column j, then q_{i_2} will be a subset of q_{i_1} which is not possible under our assumptions.

Hence all the 1's and in particular the first 1 in the row corresponding to q_{i_2} appear after the column j. Now the first 1 in the row corresponding to q_{i_k} appears on or before the j^{th} column. This is

$$\begin{array}{cc} & j \\ q_{i_1} & 1 \\ q_{i_2} & 0 \\ q_{i_3} & 0 \\ . & . \\ . & . \\ . & . \\ q_{i_k} & 1 \end{array}$$

Figure 10: A subsequence of some column of M

contradictory to our hypothesis that the queries are ordered as per the appearance of first 1. The same argument holds for $q_{i_3}, q_{i_4}, \dots q_{i_{k-1}}$. Thus such a sequence cannot exist in any column. So, M has the consecutive 1's property for columns also. $\square$

6 Concluding Remarks

If a given set of queries and records does not have the CRP, then we are interested in finding organizations where records can be stored more than once in order to ensure that all the records pertinent to a query are stored consecutively. The redundancy of such an organization is defined to be the length of such an organization minus the number of records. Obviously the problem is to find an organization with minimum redundancy. This problem is known as the *Consecutive Retrieval With Minimum Redundancy* (CRWMR) problem. The CRWMR problem is NP-complete [2]. Therefore the focus is on polynomial time approximation algorithms. In [2], heuristic algorithms for the same are presented. The only disadvantage of these algorithms is that there is no provable bound on the amount of redundancy and the effectiveness of the algorithm is only experimentally evaluated. To the best of our knowledge, this problem has not been studied from a graph theoretic point of view. We suggest the development of efficient polynomial time approximation algorithm with a provable bound on redundancy as a direction for further research.

Acknowledgments

The authors like to thank Terry McKee for several helpful discussions, an anonymous referee for his help in improving the proof of Theorem 3.1, and the Managing Editor David Gries for his comments and suggestions.

References

[1] K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *Journal of Computer and System Sciences*, 13:335–379, 1976.

[2] J. S. Deogun, V. V. Raghavan, and T. K. W. Tsou. Organisation of clustered files for consecutive retrieval. *ACM Trans. on Database Systems*, 9(4):646–671, 1984.

[3] K. P. Easwaran. Placement of records in a file and file allocation in a computer network. *Inform. Processing*, pages 304–307, 1974.

[4] K. P. Easwaran. Faithful representation of a family of sets by a set of intervals. *SIAM J. of Computing*, 4(1):56–68, 1975.

[5] D. R. Fulkerson and O. A. Gross. Incidence matrices and interval graphs. *Pacific J. Math.*, 15(3):835–855, 1965.

[6] S. P. Ghosh. File organisation: The consecutive retrieval property. *Commun. ACM*, 15(9):802–808, 1972.

[7] S. P. Ghosh. On the theory of consecutive storage of relevant records. *J. Inform. Sci.*, 6(1):1–9, 1973.

[8] S. P. Ghosh. *Database Organisation for Data Management*. Academic Press Inc., 1977.

[9] M. C. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press Inc., 1980.

[10] G. S. Lueker. *Efficient algorithms for chordal graphs and interval graphs*. PhD thesis, Princeton University, Program in Applied Mathematics and the Department of Electrical Engineering, Princeton, New Jersey, 1975.

[11] F. Roberts. Indifference Graphs. In F. Harary, editor, *Proof Techniques in Graph Theory*, pages 139–146. Academic Press, New York, 1969.

[12] D. J. Rose, R. E. Tarjan, and G. S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Computing*, 5(2):266–283, 1976.