

Implementing and Applying Dominant Graph for Dominant Relationship Analysis

Sandesh Dhawale¹, Vijay Ghorpade²

P.G. Student, Department of Computer Science and Engineering, Dr D Y Patil College of Engineering and Technology, Kolhapur, Maharashtra, India¹

Guide, Dr D Y Patil College of Engineering and Technology, Kolhapur, Maharashtra, India²

ABSTRACT: Dominant relationship concept is used for answering preference queries. The concept of dominant relationship has been extended here for the purpose of doing business analysis. New kind of analysis presented here which is called as dominant relationship analysis. This type of analysis can help product manufacturing companies to design new products, compare company's products with products of competitor company and to find interesting subspaces of the product which can be helpful for them for promoting the products. Algorithms which use Dominant Graph(DG) to answer dominant relationship queries are presented here.

KEYWORDS: dominant relationship analysis, dominant graph, dominant relationship queries.

I. INTRODUCTION

Dominant relationship concept has recently attracted much interest to answer preference queries. Here, the concept is extended for the purpose of doing business analysis using DG. If there are two records p and q in a multidimensional space, we can say that p dominates q if following two conditions are satisfied 1) in any dimension I_i , the value of p is larger than or equal to q , i.e. $p.I_i \geq q.I_i$ 2) there must exist at least one dimension I_j , such that $p.I_j > q.I_j$.

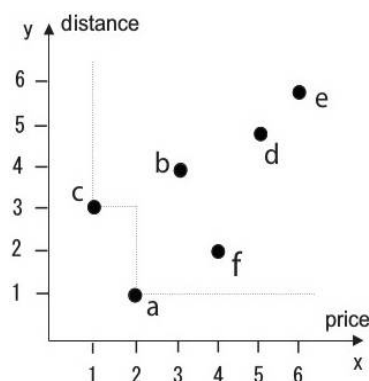


Figure 1: Example of skyline query

Preference queries are used to find those points, which are not dominated by others, in a n -dimensional dataset. Fig. 1 shows example of preference query where customers are always interested in hotels which are better in minimum in one of two dimensions, the distance and the price as compared to other hotels. In example shown in figure 1 a and c

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

are products which are not dominated by any other products. Like dominance concept is useful for customer to find products which they like, it can also be useful for manufacturing companies to determine that how popular are their products among customers in comparison to their competitor company's product. Suppose that you are a manager of hotel company. You want to know the position of hotel b in market with respect to preference like hotel's price and hotel's distance from the beach by finding number of hotels that are better or worst than hotel b. From the fig 1, it can be easily identified that hotel b is better than 2 other hotels but worse than another 2 hotels with respect to preference given. Here, five types of queries are presented for dominant relationship analysis. Queries are as follows (i) Linear Optimization Queries (LOQs), (ii) Subspace Analysis Queries (SAQs), (iii) Comparative Dominant Queries (CDQs), (iv) Skyline Product Query (SPQ) and (v) Skyline Subspace Query. This dominant relationship analysis can help product manufacturing companies to design new products, to compare company's products with competitor company's products and to find subspaces of a product which can be helpful for promoting the products. Algorithms are designed for answering these queries using DG. The paper is organized as follows: In section 2 the preliminaries of this paper are presented. Section 3 discusses related work. Section 4 defines problem statement. In section 5 algorithm for computation of DG is presented. Section 6 presents algorithms for dominant relationship queries. Conclusion is presented in section 7.

II. PRELIMINARIES

This section defines some terms used in this paper and also makes some assumptions. After that five types of dominant relationship queries are presented which are used for analysis.

Assumptions

There is an assumption made here that there are two manufacturing companies A and B, they produce a set of products $PA = \{A_1, \dots, A_s\}$ and $PB = \{B_1, \dots, B_t\}$, respectively. There is also a set of customer preferences $C = \{C_1, \dots, C_n\}$. Products or customer preferences can be represented by a point in an N -dimensional space, D , with dimensions D_1, \dots, D_N are the attributes of the products and customer preferences. It is also assumed here that product s or customer preferences are having minimum attributes.

Definitions

dominating(p, C, D'):- If an object p , a set of objects C and a set of dimensions $D' \subseteq D$ are given, then dominating(p, C, D') is defined as the set of objects from C which are dominated by p in the subspace D' of D .

dominated(C, p, D'):- If an object p , a set of objects C and a set of dimensions $D' \subseteq D$ are given, then dominated(C, p, D') is defined as the set of objects from C which dominate p in the subspace D' of D .

Dominant Relationship Queries

Linear Optimization Query (LOQ) : Given a plane L , and a set of objects, C , in an N -dimensional space of D , we define $LOQ(L, C, D)$ as the aggregate $\max(|\text{dominating}(p, C, D)|)$, such that p is any point in the plane L . This query can product manufacturing companies in designing new products that satisfy most of the customer preferences while remaining profitable.

Subspace Analysis Query (SAQ) : Given a set of points C and a point p in the N -dimensional space of D' , SAQ finds: 1. $|\text{dominating}(p, C, D')|$ and 2. $|\text{dominated}(C, p, D')|$ where $D' \subseteq D$. Products have different attributes like a smart phone have attributes like processor speed, ram size, internal memory, camera, battery, price. When all attributes of a product are considered then it may not satisfy many customer preferences but if we consider few attributes of the product then same product may satisfy more number of customer preferences. By using subspace analysis query it will

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

be easier for product manufacturing company to identify subspaces of a product where product dominates more customer preferences in the selected subspace.

Comparative Dominant Query (CDQ): Since there is tremendous competition among product manufacturing companies these days, these companies may be interested in comparing their own products with products of competitor company, doing so that they will be able to assess the position of their products in market. CDQ can help companies to achieve this goal. CDQ helps product manufacturing companies to identify customer preferences which are dominated by products of both a company and its competitor company. CDQ also helps in identifying customer preferences that are dominated by a company's products and not by competitor company's products. CDQ is defined as follows. $g_{\text{dominating}}(A, C, D)$:- If set of objects A and C in an N -dimensional space of D are given, then $g_{\text{dominating}}(A, C, D)$ is defined as the set of objects in C which are dominated by some object from A. $CDQ - (A, B, C, D)$:- If three sets of objects A, B, C in the N -dimensional space of D are given, then $CDQ - (A, B, C, D)$ is defined as: $|g_{\text{dominating}}(A, C, D) - g_{\text{dominating}}(B, C, D)|$. Similarly if three sets of objects A,B,C in the N -dimensional space of D are given, then $CDQ \cap (A, B, C, D)$ is defined as as: $|g_{\text{dominating}}(A, C, D) \cap g_{\text{dominating}}(B, C, D)|$.

Skyline Product Query (SPQ): SPQ is useful for product manufacturing companies in designing new products which are not dominated by any existing product in market. Given a set T_e of existing products in market, SPQ creates a set of best possible products from source tables T_1, T_2, \dots, T_n of sub-products such that the newly created products are not dominated by any existing products.

Skyline Subspace Query (SSQ): SSQ can help product manufacturing companies to find subspaces of a product where the given product is not dominated by any customer preferences or any existing product in market. Given a set C of customer preferences and a point p in the N -dimensional space D. Subspace Sky-line Query can be defined as determining all possible subspaces where point p is not dominated by any customer preference.

III. RELATED WORK

Many techniques have been proposed for dominant relationship analysis. Here relevant work is described. Work is divided into three categories: data mining methods for market analysis, dominant graph method, skyline query method.

Data Mining Techniques for Business Analysis

This work is motivated from the work in [1] where authors say that interestingness of knowledge being discovered depends on its usefulness to the organization. Examples in [1] discusses utility oriented mining which includes sensitivity analysis, market segmentation. Work in [2] includes profit oriented association rule discovery for product assortment. [3] discusses customer oriented catalog segmentation problem. [4] discusses profit oriented pattern discovery. [5] discusses maximum-profit item selection with cross-selling consideration. [6] discusses data mining as sensitivity analysis. In [7] dominant relationship analysis model is presented.

Dominant Graph

The DG index was proposed in [8]. In dominant graph records are linked according to dominant relationship. Based on off-line built the DG, Traveler algorithm has proposed to answer a top-k query.

V. DOMINANT GRAPH

Defining DG

Set S of records in a multi-dimension space is given, S does have k nonempty maximal layers $L_i, i=1 \dots n$. The records r in i th layer and records r' in (i+1)th layer form a bipartite graph $g_i, i=1 \dots (m-1)$. Directed edge will be there from r to r' in g_i if and only if record r dominates r' . This directed edge represents "parent-children relationship". All bipartite graphs g_i are joined to obtain DG. The maximal layer L_i is called i th layer of DG. There are two types

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

of dominant graphs dominated dominant graph and dominating dominant graph. Dominated dominant graph shows dominated relationship among records in graph. Dominating dominant graph shows dominating relationship among records in graph.

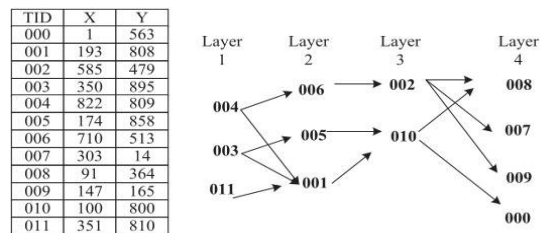


Figure 2: Dominant Graph for given database

Figure 2 shows a sample dataset consisting of 11 records and DG built for it using algorithm described below

Algorithm

Input: i) Dominant graph of the database. ii) r = record to be inserted.

- 1) If r is not dominated by any record in first layer of DG then
- 2) set $n=0$
- 3) else
- 4) All record P_i at first layer of DG that dominates r are collected to form the set P .
- 5) Do DFS search from each P_i to find the longest path L , and each record in dominates r .
- 6) Set $n = |L|$.
- 7) Insert r into the $(n+1)$ th layer of DG.
- 8) If r dominates some records C_i in the $(n+1)$ th layer of DG then
- 9) All the descendents records of C_i (including C_i) are collected to form the set S .
- 10) For each record O in S do
- 11) O is degraded into it's next layer.
- 12) Build parent children relationship between O and records in current next layer.
- 13) if O has some other parrent A that is not in set S then
- 14) Delete the directed edge from A to O .
- 15) Build parent children relationship between records in n th layer and r .
- 16) Build parent children relationship between records in r and $n+2$ th layer.
- 17) Report the updated DG.

IV. PROBLEM STATEMENT

A Dominant Relationship Analysis Tool is presented here for product manufacturing companies. This analysis tool can be used for helping product manufacturing companies to design new products and to extracts important information from large dataset. Extracted information will be very helpful to manufacturers to do the business analysis, to determine the position of their products in the market more effectively, and to identify interesting subspaces of a product which can be used for promoting the products .

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

V. ANSWERING DOMINANT RELATIONSHIP QUERIES

In this section, algorithms are proposed to answer dominant relationship queries using DG.

1. Linear Optimization Query

Given a plane L, a set of objects C in space D, we wish to find some cells which are intersecting plane L and dominate the most points in C.

Algorithm

- 1) Input: A plane L, set of objects C .
- 2) Output: Points on plane L which dominates maximum number of objects from C.
- 3) Method:
 - i) Start from the point 1,1,...1 in N-dimensional space.
 - ii) At any stage if the cell is at the bottom left of plane L, iterate continually to it's children till we find the cell on the plane L.
 - iii) If the cell is on the plane L add it to the result cell set.
 - iv) Obtain the cells from result cell set which dominate maximum number of objects from set C using LOQ Max Dominating function.
- 4) Function LOQ Max Dominating (result cell set, DG for set C)
 - i) Find the position of the cell in DG for set C.
 - ii) Determine number of objects in set C are dominated by the cell.
 - iii) Repeat step 1 and 2 for each cell in result cell set.
 - iv) Return the cells which dominate more objects in set C than any other cell in result cell set.

2. Subspace Analysis Query

If we are given a set of points C and a point p in the N-dimension of D, the most basic SAQ is to find for each subspace D' , how many points dominating or dominated by p.

Algorithm

- 1) Input: i) C = Set of objects in N-dimensional space D.
 - ii) D' = A subspace of dimension D.
 - iii) A point P.
- 2) Output: i) Number of points in C dominated by P.
 - ii) Number of points in C dominating P.
- 3) Method:
 - i) Build dominating dominant graph DG1 for set of objects in C.
 - ii) Find the position of product p in DG1 and determine how many objects form C are dominated by point p.
 - iii) Build dominated dominant graph DG2 for set of objects in C.
 - iv) Find the position of product p in DG2 and determine how many objects form C are dominating point p.

3. Comparative Dominant Query

Comparative dominant queries are those queries which are used to compare the set of dominated objects between competitive products. There are two kinds of CDQs: $CDQ-(A,B,C,D)$, which retrieves $|gdominating(A,C,D)-gdominating(B,C,D)|$; and $CDQ \cap (A,B,C,D)$, which retrieves $|gdominating(A,C,D) \cap gdominating(B,C,D)|$.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Algorithm

- 1) Input: i) Three sets of objects A, B, C.
ii) Dominant graph for C.
- 2) Output: i) Set of objects in C which are dominated by some objects in A and not by any object in B.
ii) Set of objects in C which are dominated by some objects in A and some objects in B.
- 3) Method:
 - i) Take an object x from set A.
 - ii) Find the position of that object x in dominant graph.
 - iii) Determine objects in C which are dominated by the object x.
 - v) Repeat steps 1, 2, 3 for each object in set A.
 - v) Let S1 be the set of objects in C which are dominated by objects in A.
 - vi) Take an object y from set B.
 - vii) Find the position of the object y in dominant graph.
 - viii) Determine objects in C which are dominated by the object y.
 - ix) Repeat steps 6, 7, 8 for each object in set B.
 - x) Let S2 be the set of objects in C which are dominated by objects in B.
 - xi) $CDQ(A, B, C, D) = S1 - S2$.
 - xii) $CDQ \cap (A, B, C, D) = S1 \cap S2$.

4. Skyline Product Query

Given a set T_e of existing products in market, a set of best possible products is to be created from source tables T_1, T_2, \dots, T_n of sub-products such that the newly created products are not dominated by any existing products in T_e .

Algorithm

- 1) Input: i) Dominant graph for existing products.
ii) Source tables of sub products T_1, T_2, \dots, T_n .
- 2) Output: Table of newly created skyline products.
- 3) Method:
 - i) Build dominant graph for existing product dataset T_e .
 - ii) Generate new product table T_p by combining subproducts in tables $T_1', T_2' \dots T_k'$.
 - iii) Find a set T_p' of all products from T_p which dominate all the products in skyline of DG of T_e .
 - vi) Return T_p' .

5. Skyline Subspace Query

Given a set C of customer preferences and a point p in the N -dimensional space D. Subspace Sky-line Query can be defined as determining all possible subspaces where point p is not dominated by any customer preference.

Algorithm

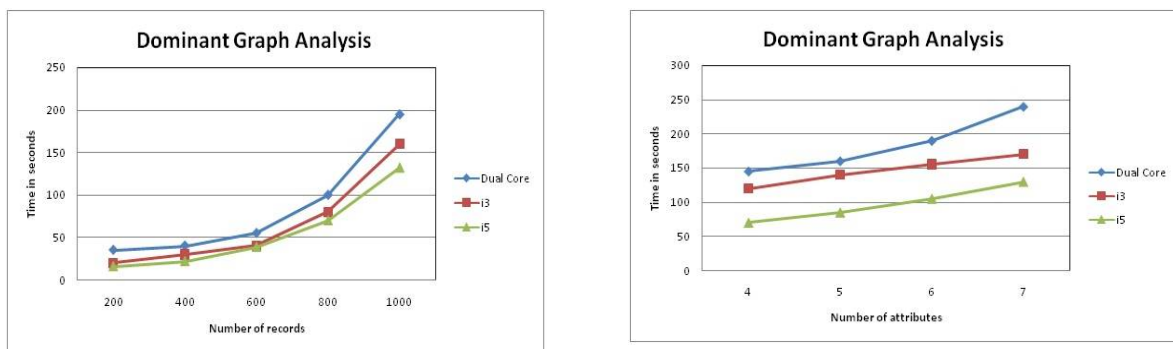
- 1) Input: i) C: An N dimensions set of customer preference
ii) A point p.
- 2) Output: All the subspaces of product P where p is in skyline.
- 3) Method:
 - i) Create DG for each dimension in D for dataset C.
 - ii) Read point p.
 - ii) Obtain the sets S_1, S_2, \dots, S_n such that each point in S_i dominates p in the dimension D_i
 - iv) if $S_1 \cap S_2 \dots \cap S_k = \text{NULL}$ then Point p is a skyline point in subspace $\{D_1, D_2, \dots, D_k\}$.
 - vi) For all combinations of dimensions in D check whether p is in skyline or not.

VI. EXPERIMENTAL SET UP AND RESULT ANALYSIS

To evaluate the efficiency and effectiveness of dominant graph and dominant relationship queries experiments were conducted. All algorithms implemented using Neatbeans 8.2 and oracle 10g database, and conducted experiments on PCs with different configurations such as dual core processor with 2GB memory, i3 processor with 3GB memory and i5 processor with 8GB memory. Experiments conducted on operating systems like Windows 7 and Ubuntu 14.04

1. Efficiency of DG Computation

Running DG algorithm on PCs with different configurations



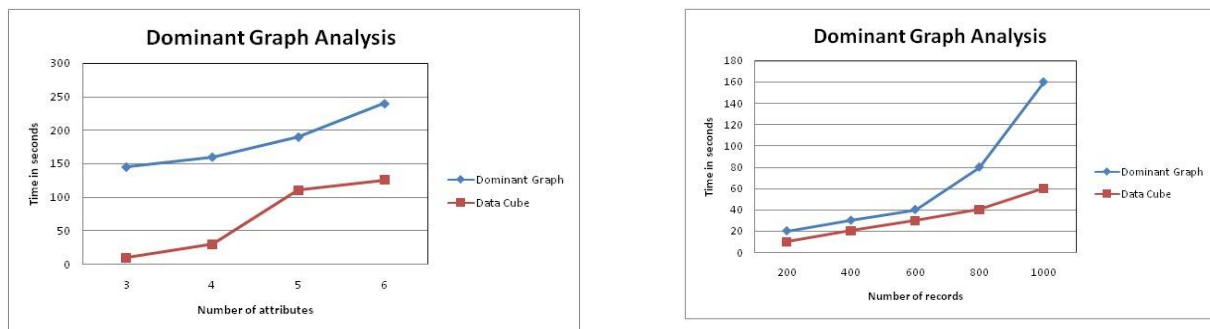
(a) Varying records

(b) Varying attributes

Figure 3: Dominant graph analysis using PC's of different configuration

Here to analyze the performance of the algorithm for computing dominant graph, we ran the algorithm on PCs of different configurations such as dual core processor with 2GB memory, i3 processor with 3GB memory and i5 processor with 8GB memory. All PCs are running on same operating system i.e. Windows 7. Figure 5.1-a shows run time for algorithm when number of attributes are fixed to 7 and number of records are varied from 200 to 1000. Figure 5.1-b shows run time for algorithm when number of records fixed to 1000 and number of attributes varied form 4 to 7.

Dominant graph Vs Data cube



(a) Varying records

(b) Varying attributes

Figure 4: Dominant graph and Data cube comparison

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

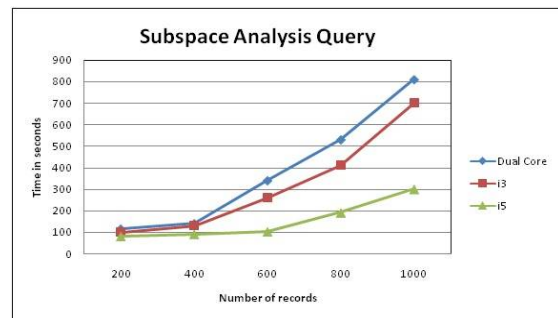
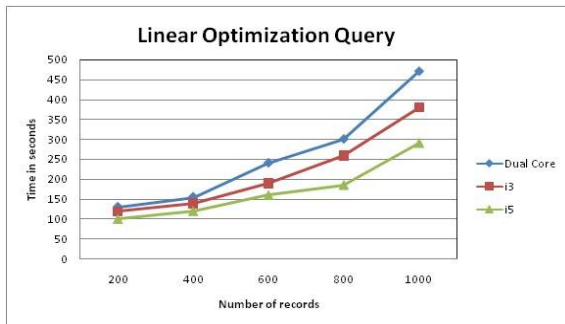
Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

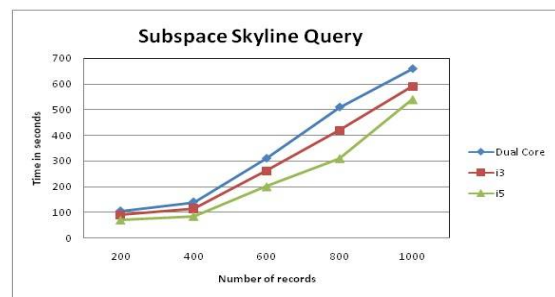
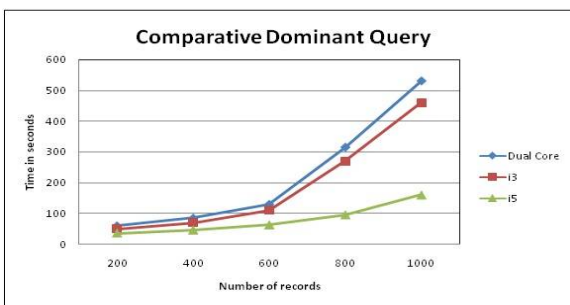
Here we compare performance of algorithm for computing dominant graph and algorithm for computing data cube. Experiment is conducted on PC with Pentium IV processor and 2GB memory. Figure 5.3-a shows run time for algorithm when number of attributes are fixed to 7 and number of records are varied from 200 to 1000. Figure 5.3-b shows run time for algorithm when number of records fixed to 1000 and number of attributes varied form 4 to 7.

2. Query Answering Performance

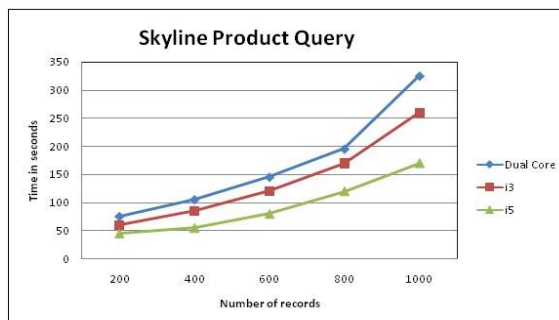
Measuring runtime performance of dominant relationship queries by running them on PCs with different configurations



Here Linear Optimization Query and Subspace Analysis Query are analyzed by comparing their runtime on PCs with different hardware configuration.



Here Comparative Dominant Query and Subspace Skyline Query are analyzed by comparing their runtime on PCs with different hardware configuration.



International Journal of Innovative Research in Science, Engineering and Technology

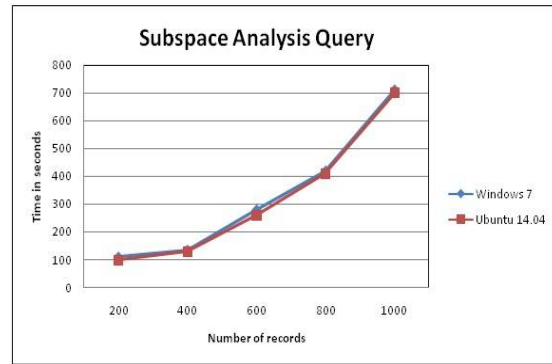
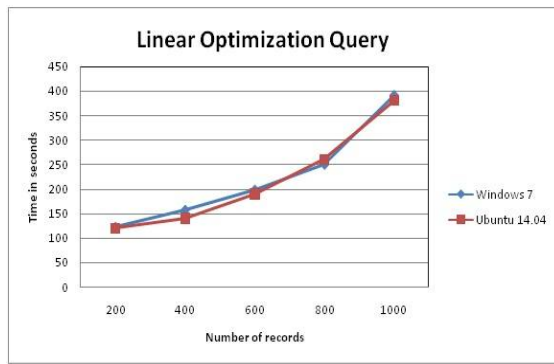
(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

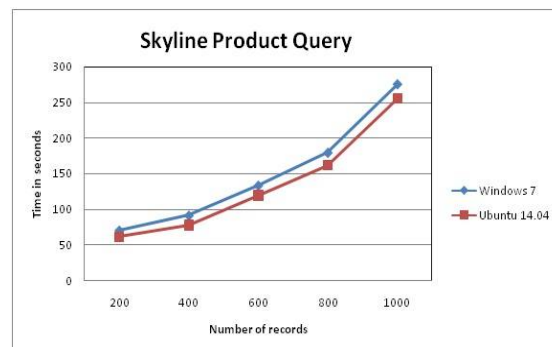
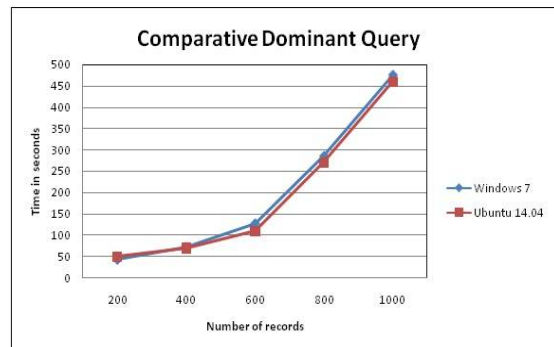
Vol. 6, Issue 5, May 2017

Here Skyline Product Query is analyzed by comparing it's runtime on PCs with different hardware configuration. Here while running algorithms for queries, number of attributes are fixed to 7 and number of records varied from 200 to 1000.

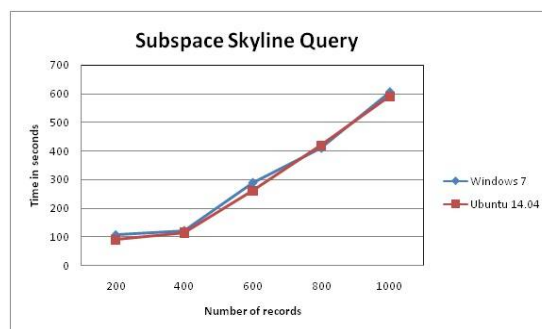
Measuring runtime performance of dominant relationship queries by running them on PCs with different operating systems



Here we analyze Linear Optimization Query and Subspace Analysis Query by running them on PC having operating systems Windows 7 and PC having operating system Ubuntu 14.04.



Here we analyze Comparative Dominant Query and Skyline Product Query by running them on PC having operating systems Windows 7 and PC having operating system Ubuntu 14.04.



Here we analyze Subspace Skyline Query by running it on PC having operating systems Windows 7 and PC having operating system Ubuntu 14.04.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Here while running algorithms for queries, number of attributes are fixed to 7 and number of records varied from 200 to 1000.

VII. CONCLUSION

The analysis tool developed here is for product manufacturing companies to increase their business. Analysis tool can help manufacturing companies to create new profitable products which satisfy most of the customer preferences, to compare company's products with competitor company's product by checking how many of them satisfy customers requirement and to find subset of product's attributes that satisfy most of the customers requirements. This analysis tool consist of five types of queries and these queries are implemented using dominant graph.

From experiments conducted it can be noticed that the performance of algorithm for computing dominant graph is good when it is run on PC with descent configuration.

From experiments it can be observed that dominant graph computation is bit slower than data cube computation, hence there is scope to improve the efficiency of the algorithm which is used for computing dominant graph in future.

From experimental study we can also see that dominant relationship queries take bit less time for execution when they are run on Ubuntu 14.04 operating system compared to Windows 7. Dominat relationship queries also need PC with descent configuration, so there is scope to improve the efficiency of algorithms for dominant relationship queries.

REFERENCES

- [1] Kleinberg J, Papadimitriou C, Raghavan P "A microeconomic view of data mining", Data Min Knowl Discov 2(4):311-322, 1998.
- [2] Brijs T, Swinnen G, Vanhoof K, Wets G , "Using association rules for product assortment decisions: a case study", ACM SIGKDD international conference on Knowledge discovery and data mining, pp 254-260,1999.
- [3] Ester M, Ge R, Jin W, Hu Z, "A microeconomic data mining problem: customer-oriented catalog segmentation", ACM SIGKDD international conference on knowledge discovery and data mining, pp 557-562, 2004.
- [4] Wang K, Zhou S, Han J, "Profit mining: from patterns to actions", International conference on extending database technology, pp 70-87, 2002.
- [5] Wong R, Fu A, Wang K, "MPIS: maximal-profit item selection with cross-selling considerations", IEEE international conference on data mining, pp 371-378, 2003.
- [6] Yao J "Sensitivity analysis for data Mining", International conference of the North American fuzzy information processing society, pp 272-277, 2003.
- [7] C. Li, B.C. Ooi, A.K.H. Tung and S. Wang. DADA: A data cube for dominant relationship analysis. ACM SIGMOD International Conference on Management of Data, pp. 659-670, 2006.
- [8] Lei Zou, Lei Chen, "Pareto-Based Dominant Graph: An Efficient Indexing Structure to Answer Top-K Queries", IEEE transactions on Knowledge and Data Engineering, Vol.23,5, 2011.