ISSN:0975-9646

Enhancing Query Efficiency Using Pruning Techniques on Incomplete Data

Dr.D.Bujji Babu[#1], G.Swapna[*2]

[#]*MCA Department, JNTK University*
QIS College of Engg&Technology,

***Abstract*— The Top-k Dominating query returns the k data objects which dominate the highest number objects in a data set. This query is an important tool for Decision support since, It provides data analysis an institutive way for finding significant objects .In addition,It combines the Advantages of top-k and skyline queries without sharing their Disadvantages:i)The output size can be controlled ii) no ranking functions need to be specified by users, and iii)the results independent of the scales at different dimensions despite their importance ,top-k dominating queries have not received adequate attention from the research community .In this paper we design specialized algorithms that apply on indexed multi-dimensional data and fully exploit the Charactrestics of the problem.experiments on synthetic dataset demonstrate that our algorithms significantly outperform a previous sky-line based approach, while our results on real datasets show the meaningfulness of top-k dominating queries**

***Keywords*— Movie Ratings, Multi Dimensional Data,progressive algorithms,IncompleteData ,Dominance Relation**

INTRODUCTION:

Top-k query processing has an active research area spanning like search, p2p-based retrieval,multimedia databases,to name a few.The query's power lies in its flexibility,supporting user-defined and ad-hoc scoring functions,and its ability to bound the number of results through the parameter k.Unfortuntely,the selection of a meaningful scoring function is not always easy,since different scoring functions is not generally produced different results.Moreover,top-k queries are sensitive to value scaling.

Due to the importance of skyline queries,several research efforts have been dedicated to develop efficient skyline query processors[2],[3]..Almost all of these algorithms rely mainly on two implicit assumptions:First assumption is Data are complete,i.e.,all dimensions are available for all data items.Such an assumption of completeness is not practical in many users. For,Consider the movie rating application[1] with hundreds of users rating thousands of movies.It is highly unlikely that every single user will rate all movies.Instead, a user will rate only the movies that interest her. As a result ,Each movie will be represented as a D-dimensional point with several blank(i.e., incomplete)dimensions.Second assumption is with the exception of all skyline algorithms assume transitivity in the dominance relation,i.e if data item p_iDominates p_j while p_j dominates $p_{K,}$ then p_i dominates p_k .Using the transitivity property ,skyline query processing algorithms exploit various ways of data pruning and indexing Unfortunately .as will be seen in this paper,the transitive dominance relation is not applicable to the case of incomplete data

For ease of representation and computation ,we represent a D-dimensional incomplete point p by a bitmap vector P.B of D bits that include 1's for all complete dimensions and 0's for all incomplete dimensions P=(4,-5,-) and Q=(-3,3,2) are represented by the bitmaps P.B=1010 and Q.B=0111,respectively

A naïve solution for incomplete data is to do an exhaustive pairwise comparisons between all input points and select only those points that are not dominated.For very large input sizes,this naïve solution is not feasible.In this section,we improve upon the naïve solution by introducing two new algorithms,namely ,the replacement and the bucket algorithms that tailor existing skyline algorithms to work for incomplete data

In this paper,we go beyond the completeness assumption of multi-dimensional data where we develop new algorithms for efficient computation of skyline queries over incomplete data sets

LITERATURE SURVEY:

A literature survey or a literature review in a project report is that section which shows the various analyses and research made in the field of your interest and the results already published, taking into account the various parameters of the project and the extent of the project.

Recently, there has been much interest in processing skyline queries for various applications that include decision making, personalized services, and search pruning. Skyline queries aim to prune a search space of large numbers of multidimensional data items to a small set of interesting items by eliminating items that are dominated by others. Existing skyline algorithms assume that all dimensions are available for all data items. This paper goes beyond this restrictive assumption as we address the more practical case of involving incomplete data items (i.e., data items missing values in some of their dimensions). In contrast to the case of complete data where the dominance relation is transitive, incomplete data suffer from non-transitive dominance relation which may lead to a cyclic dominance behavior. We first propose two algorithms, namely, "Replacement" and "Bucket" that use traditional

skyline algorithms for incomplete data. Then, we propose the "ISkyline" algorithm that is designed specifically for the case of incomplete data. The "ISkyline" algorithm employs two optimization techniques, namely, virtual points and shadow skylines to tolerate cyclic dominance relations. Experimental evidence shows that the "ISkyline" algorithm significantly outperforms variations of traditional skyline algorithms

This paper has addressed the problem of skyline queries over incomplete data where multi-dimensional data items are missing some values of their dimensions. We showed that with incomplete data, the dominance relation among data points may not be transitive, thus, almost all existing techniques for skyline queries are not applicable. We have proposed two new algorithms, namely, the Replacement and the Bucket algorithms that utilize variations of traditional skyline algorithms to accommodate incomplete data. Then, we proposed the ISkyline algorithm that is designed specifically for incomplete data. The ISKyline algorithm employs two optimization techniques, namely virtual points and shadow skylines to exploit the properties of incomplete. The correctness of the ISkyline is proved in terms that produce only and all skyline points. Experimental results based on real and synthetic data sets show the efficiency and scalability of the ISkyline algorithm.

METHODOLOGY:

Algorithms:

A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

1.Extended Skyband Based Algorithm (ESB):

Input: an incomplete data set S, a parameter k
Output: the result set SG of a TKD query on S
/*kSB(O): the result set of a k-skyband query on a bucket O */
1: initialize sets $S_C \leftarrow S_G \leftarrow \phi$
2: for each object o ЄS do
3: insert o into a bucket O based on $_{\beta 0}$ (create O if necessary)
4: for each bucket O do
5: $S_C \leftarrow S_C \bar{U}$ KSB(O)
6: for each object o Є S_C do
7: update score(o) by comparing o with all the objects in S
8: add the k objects in SC having the highest scores to SG
9: return S_G

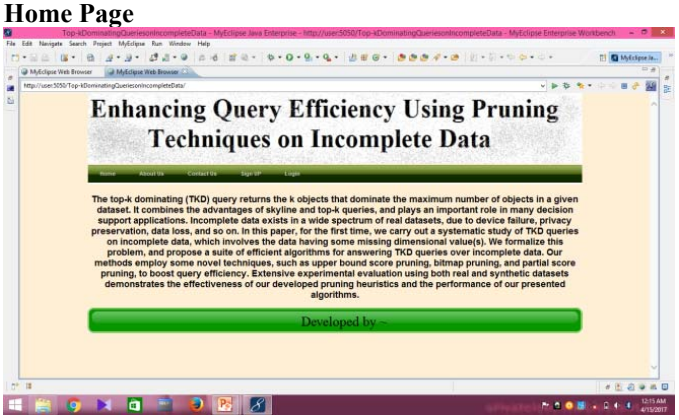
Upper Bound Based Algorithm:

Input: an incomplete data set S, a parameter k, a pre-computed
priority queue F sorting all objects from S in descending order of their MaxScore
Output: the result set S_G of a TKD query on S
1: initialize sets $S_C \leftarrow S_G \leftarrow \phi$ and $\tau \leftarrow -1$
2: while F is not empty do
3: o ← de-queue(F)
4: if Max Score (o) ≤ τ then break //Heuristic1
5: else
6: score(o) Get _ Score(o)
7: if score(o) > τ or τ < 0 then
8: $S_C \leftarrow S_C \bar{U} \{o\}$
9: if $|S_C| > k$ then
$S_C \leftarrow S_C - \{p\}$ with p ЄS_C and score(p)= τ
11: update τ min {score(c) c Є S_C} if |SC|=k
12: return $S_G \leftarrow S_C$

Screen Shots:

A Screen shot is an image taken by a person to record the visible items displayed on the moniter.It is usually,this is a digital image using the operating system

Home Page

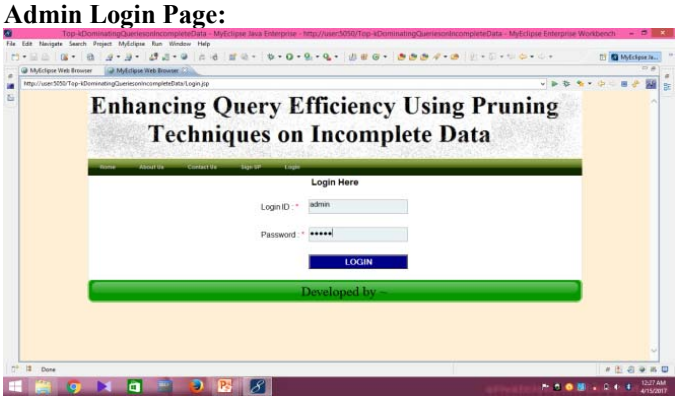


Screen 1:Home Page

Description:

In The above screen is home page .it contains the home ,about us ,contact us,sign up,login fields etc.and after user register the all details then click the login option

Admin Login Page:

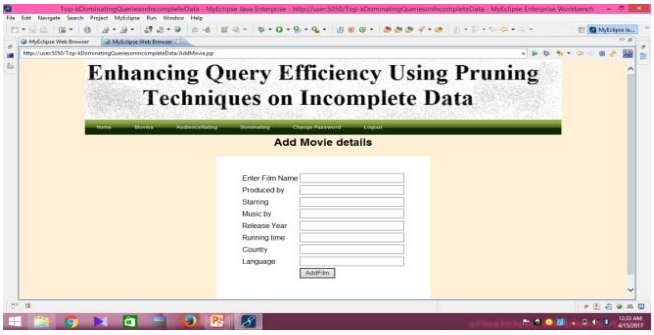


Screen 2:Admin Login Page

Description:

The above screen is the admin login form to be create login form two textboxes is created to enter user name and password to click the the login button then displays a admin home page.

Add Movie Details:

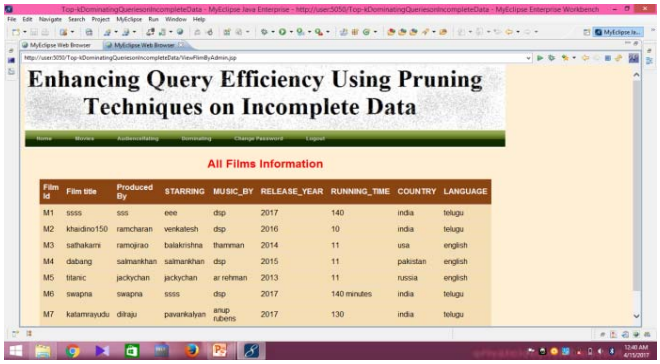


Screen 3:Add Movie Details

Description:

In The above screen is the add movie details.first we have to the open the admin home page then click the movies option after the two options are displayed .one is add movies To enter the film names,produced by,starring,music by,Release year,country,language after the click the add film button

All Films Information:

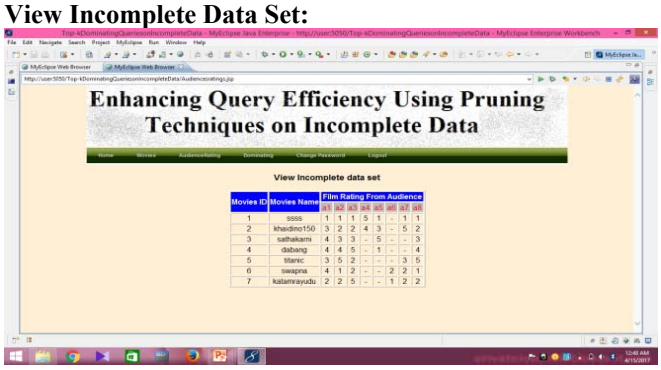


Screen 4:All Films Information

Description:

In the above screen is display the all films information to the add films

View Incomplete Data Set:

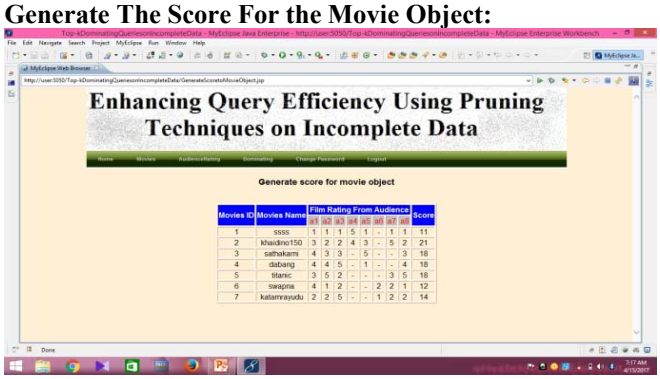


Screen 5:View Incomplete Data Set

Description:

In the above screen is display the view incomplete data set .First audiences give the rating in all films some audiences give the no rating then also form the incomplete data set

Generate The Score For the Movie Object:

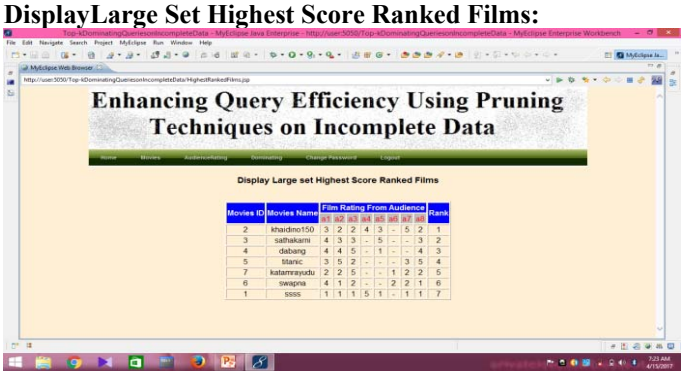


Screen 6:Generate The Score For Movie Object

Description:

In the above screen is Generate the score for the movie object.first display audience rating.then after caluculate the score for the all audiences. total score for the movie objects

DisplayLarge Set Highest Score Ranked Films:

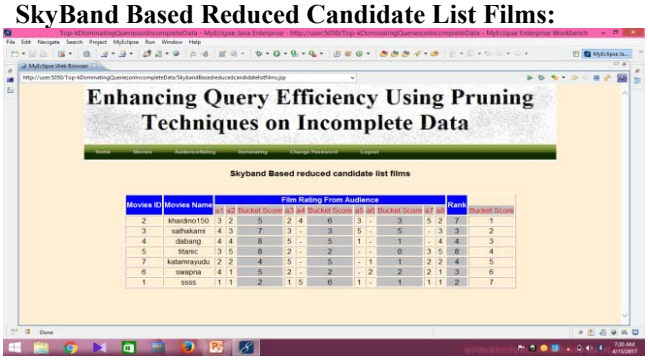


Screen 7:Display Large Set Highest Score Ranked Films

Description:

In the above screen is Display Large Set Highest Ranked Films.In the incomplete data from the audience rating.then calculate the scoe for movie objects display the highest score for the movie then the movie give the ranking purpose

SkyBand Based Reduced Candidate List Films:

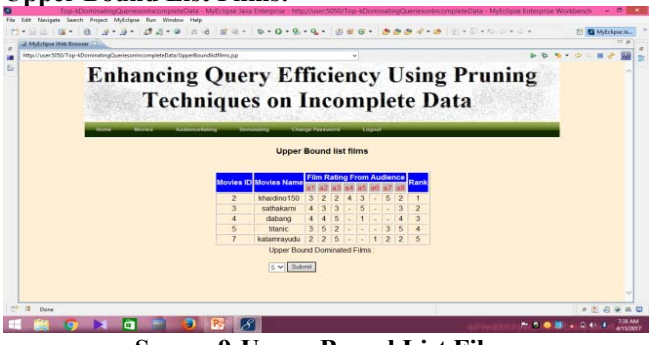


Screen 8:Skyband Based Reduced Candidate List Films

Description:

In the above screen is Skyband based candidate list films.The skyband is used to the reduced the all candidate rating films .after the reducing calculate the score for the reducing films after give the ranking so as well as skyband is also calculate the bucket score.the bucket score depends on the ranking for highest films

Upper Bound List Films:

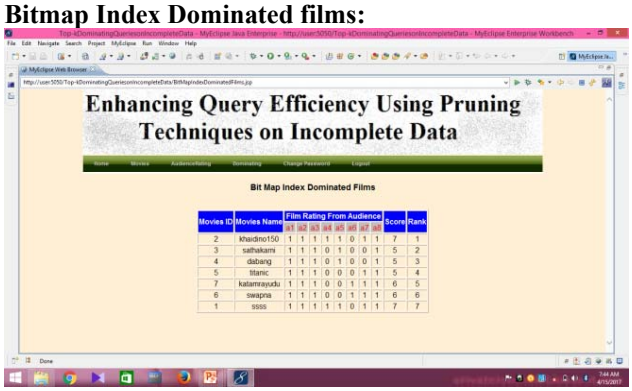


Screen 9:Upper Bound List Films

Description:

In the above screen is Upper bound list films.upper bound is also used to the display the most top of the dominated films at audience rating then after type the dominted films after click the submit button .

Bitmap Index Dominated films:

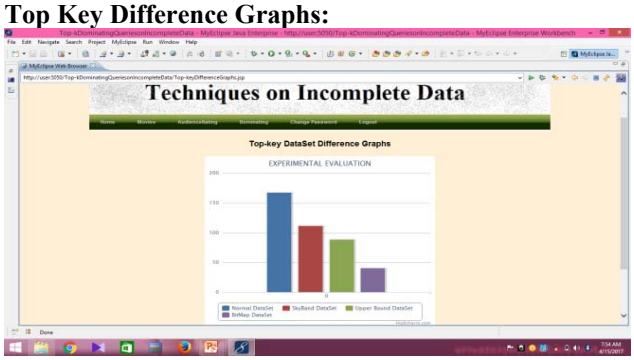


Screen 10:Bitmap Index Dominated Films

Description:

The above screen is the bitmap index dominated films.in bitmap index is used to the 0's and 1's .incomplete data represents the '0' and remaining the values are the represents the '1' after calculating the score then display the ranking score.

Top Key Difference Graphs:



Screen 11:Top Key Difference Graphs

Description:

In the above screen is top key difference graphs.the ouput of the top key difference representation in graphical representation.the graphical representation is overall the audience rating

CONCLUSIONS

Consider the wide range of applications for top-k Dominating (TKD) queries and the pervasiveness TKD query on incomplete data, we, in this paper, study the problem of the TKD query on incomplete data where some dimensional values are missing. To efficiently address this, when first propose ESB and UBB algorithms, which utilize novel techniques (i.e., local sky band technique and upupper bound score pruning) to prune the search space. In or In order to further reduce the cost of score computation, we present BIG algorithm, which employs the upper bound score pruning, the bitmap pruning and fast bitwise operations based on the bitmap index to improve the score computation and boost query performance accordingly. Moreover, in order to trade efficiency for space, we propose IBIG algorithm by using the bitmap compression technique and the binning strategy over BIG, and develop a method to choose the appropriate number of bins. Considerable experimental results on both real and synthetic datasets confirm the effectiveness and efficiency of our presented heuristics and algorithms. In the future, we will further study how to improve the quality of TKD query over incomplete data.

ACKNOWLEDGMENT

We acknowledge our sincere thanks and deep sense of gratitude to Mr.N.S.Kalyan Chakravarthy, Executive Chairman of QIS Educational Institutions, ongole, Andhrapradesh, for providing an excellent computational facilities, a very good learning environment in the campus with numerous journals and the digital library. We also thank Mr.N.Nageswara rao, President, SNES for his leadership and inspirational talks on importance of ethics and quality education.

REFERENCES

[1] M. E. Khalefa, M. F. Mokbel, and J. J. Levandoski, "Skyline query Processing for incomplete data," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 556–565.

[2] Y. Gao, X. Miao, H. Cui, G. Chen, and Q. Li, "Processing k-skyband,constrained skyline,and group-by skyline queries on incomplete data"Expert Syst. Appl., vol. 41, no. 10, pp. 4959–4974,2014.

[3] X. Lian and L. Chen, "Top-k dominating queries in uncertain databases,"in Proc. 12th Int. Conf. Extending Database Technol.: Adv.Database Technol., 2009, pp. 660–671.

[4] X. Lian and L. Chen, "Probabilistic top-k dominating queries in uncertain databases," Inf. Sci., vol. 226, pp. 23–46, 2013.

[5] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "Progressive skyline computation in database systems," ACM Trans. Database System.,vol. 30, no. 1, pp. 41–82, 2005.

[6] M. L. Yiu and N. Mamoulis, "Efficient processing of top-k dominating queries on multi-dimensional data," in Proc. 33rd Int. Conf.Very Large Data Bases, 2007, pp. 483–494.

[7] M. L. Yiu and N. Mamoulis, "Multi-dimensional top-k dominating queries," The Int. J. Very Large Data Bases, vol. 18, no. 3,pp. 695–718, 2009.

[8] W. Zhang, X. Lin, Y. Zhang, J. Pei, and W. Wang, "Threshold based probabilistic top-k dominating queries," The Int. J. VeryLarge Data Bases, vol. 19, no. 2, pp. 283–305, 2010.

[9] E. Tiakas, G. Valkanas, A. N. Papadopoulos, Y. Manolopoulos,and D. Gunopulos, "Metric-based top-k dominating queries," in Proc. Int. Conf. Extending Database Technol., 2014, pp. 415–426.

[10] L. Antova, C. Koch, and D. Olteanu, "From complete to incomplete information and back," in Proc.SIGMOD Int. Conf. Manage.Data, 2007, pp. 713–724.