Feature Based Retrival System for Inexplicite Queries

Ms.Sarika Sarode[1] Dr.Kalpana Metre[2]
[1,2]Department of Computer Engineering
[1,2]Savitribai Phule University, Nasik, India

Abstract— In this decade there are hugh development in data engineering. Handling of the hugh data and searching in large databases takes more time. FBR System will developed to retrieve the data fastly from the database. It allows user to reprocess the query result. It allows user to reprocess the query result. It allows user to express their uncertainty about input by specifying probability values. FBR System reduces the time required to retrieve the data. It is a scientific way to search the data in to the database.

Key words: FBR System, Inexplicite Queries

I. Introduction

The traditional DBMS provide facility to user to specify the query by using one particular format to retrieve the data from the database. The user who has knowledge of DB schema and other can able to retrieve the data from the database. Now, database has to handle large number of user who searching the information in the database of their interest. While searching the information in the database, the user is not certain sometimes. i.e. they are unsure about their query input. So to process the input which specified in the terms of probability value?

FBR System allows user to reprocess the query output.

It mainly focuses on three things: 1.It retrieve the data within less time.2.It turns the user input into the probability values.3.It allows user to reprocess the query output.Now, to understand the functionality of FBR System, consider the example, and suppose there is one large database of car.It contains the information related to the various models of car. Now see the information related to the car is stored in terms of features of Carlike as Type of Fuel used by Car, Color of car, Size of car, Macwheel etc.Now, considers Rahul who have seen one car. And now wants the detailed specification of the car. Then he enters the attribute values as shown in Fig 1.He initially enters the Color of Car. But he may not be sure about color. Then three options provided by FBR System to user like sure, not sure, preety sure etc..After the input taken from user the FBR will process that input and generates the report which contains three tables..i .e Fig 2.The middle table shows the different types of top ranked Car Models. The table in right shows FBR's novel sensitivity score ,which will able to support scientific analysis including inexplicit conditions. Rahul can continue his search even after positive identification.He can use other two tables to refine his queries.The left table present him

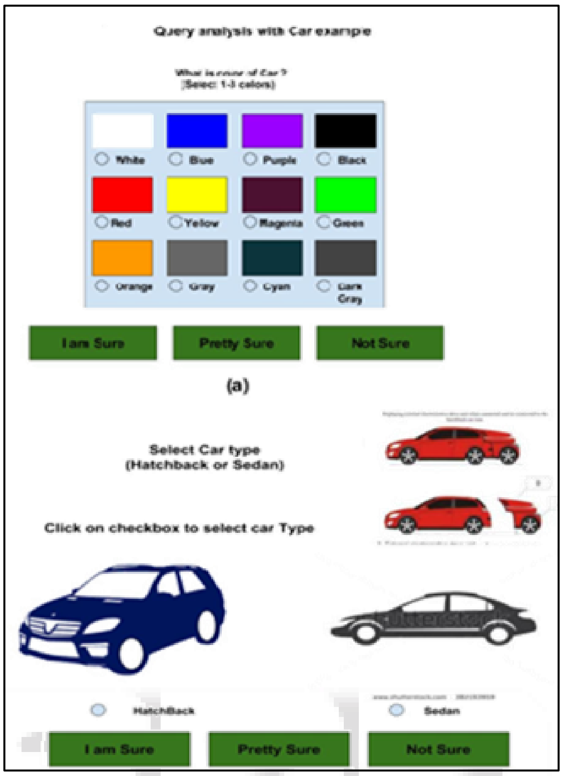
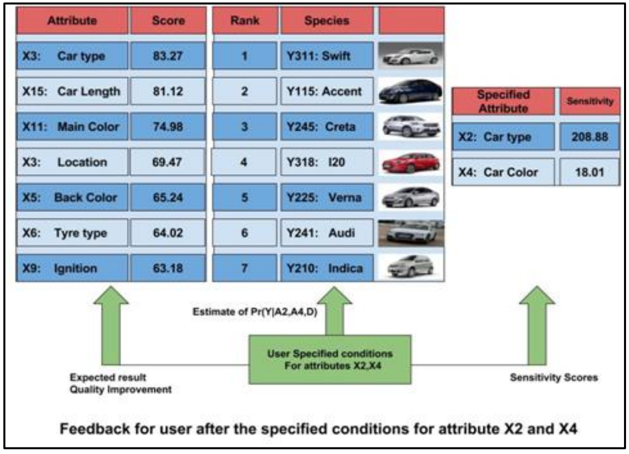


Fig. 1: Possible interfaces for specifying attributes.

the which other attribute would be most helpful in narrowing the search. System does not know that which type of Car Model user is looking for, an features potential usefulness in improving result quality is calculated on the basis of intuition that most useful features are those that that best separate "Winners" from "losers".



Attribute	Score
X3: Car type	83.27
X15: Car Length	81.12
X11: Main Color	74.98
X3: Location	69.47
X5: Back Color	65.24
X6: Tyre type	64.02
X9: Ignition	63.18

Rank	Species
1	Y311: Swift
2	Y115: Accent
3	Y245: Creta
4	Y318: I20
5	Y225: Verna
6	Y241: Audi
7	Y210: Indica

Specified Attribute	Sensitivity
X2: Car type	208.88
X4: Car Color	18.01

Estimate of Pr(Y|A2,A4,D)
User Specified conditions
For attributes X2,X4
Expected result
Quality Improvement
Sensitivity Scores
Feedback for user after the specified conditions for attribute X2 and X4

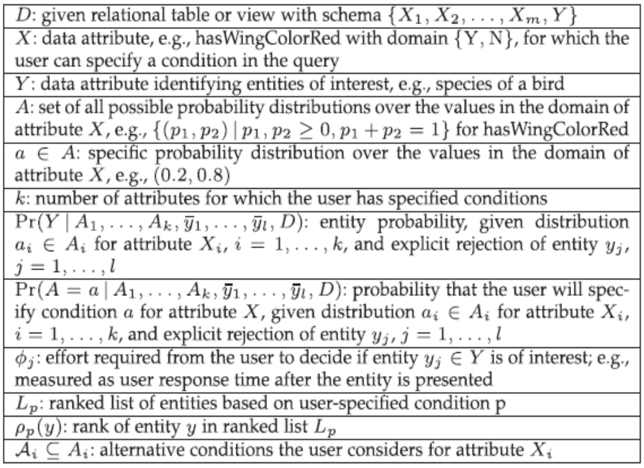
Fig. 2: Feedback to the user after user specified conditions for attributer X_2 and $X_{4.}$

For example, Bluetooth functionality will receive high score if all among Red Car's this functionality present in few of them. Whereas, if the most of the red car having Bluetooth functionality then Bluetooth functionality will receive low score.

Rahul can also determine the risk of entering a conditions by using sensitivity analyser. It measures how much the query result will change if Rahul were to alter the corresponding conditions.

II. FBR SYSTEM FRAMEWORK OVERVIEW

The data model used in FBR System propose a probabilistic framework for scientific search in databases. Important notations is summarized in figure 3.



D: given relational table or view with schema $\{X_1, X_2, \ldots, X_m, Y\}$
X: data attribute, e.g., hasWingColorRed with domain $\{Y, N\}$, for which the user can specify a condition in the query
Y: data attribute identifying entities of interest, e.g., species of a bird
A: set of all possible probability distributions over the values in the domain of attribute X, e.g., $\{(p_1, p_2) \mid p_1, p_2 \geq 0, p_1 + p_2 = 1\}$ for hasWingColorRed
$a \in A$: specific probability distribution over the values in the domain of attribute X, e.g., $(0.2, 0.8)$
k: number of attributes for which the user has specified conditions
$\Pr(Y \mid A_1, \ldots, A_k, \bar{y}_1, \ldots, \bar{y}_l, D)$: entity probability, given distribution $a_i \in A_i$ for attribute X_i, $i = 1, \ldots, k$, and explicit rejection of entity y_j, $j = 1, \ldots, l$
$\Pr(A = a \mid A_1, \ldots, A_k, \bar{y}_1, \ldots, \bar{y}_l, D)$: probability that the user will specify condition a for attribute X, given distribution $a_i \in A_i$ for attribute X_i, $i = 1, \ldots, k$, and explicit rejection of entity y_j, $j = 1, \ldots, l$
ϕ_j: effort required from the user to decide if entity $y_j \in Y$ is of interest; e.g., measured as user response time after the entity is presented
L_p: ranked list of entities based on user-specified condition p
$\rho_p(y)$: rank of entity y in ranked list L_p
$\mathcal{A}_i \subseteq A_i$: alternative conditions the user considers for attribute X_i

Fig. 3: Important Notations

A relational table or view D with schema{X_1,X_2,……X_y}.Attribute Y takes on a special role, identifying entities of interest to the user. Depending upon query any attribute of D could take on this role. For instance, in car identification example is the Model name. When looking for companies ,Y would be the corresponding company's identifier. Even though Y identifies entities of interest to the user, it does not need to be key of D.The tuples in D could represent precise or probabilistic information ,including crowd sourced inexplicit data.

In the Car's example, the entities of interest are Model of Car. The X_i describes various properties of Car .E.g. Macwheel,Bluetooth functionality etc.

Sometimes there are multiple records for One Car Model. So X_i can differ even for records with same Y-value.

A. *Exploration Framework*

Fig.3 shows the FBR system components and their interactions. From the given data set Ditto types of models are trained at "setup time", i.e. before the system is available to the user queries.The entity model and attribute model are trained by using dataset D.At runtime the entity ranker and attribute ranker.

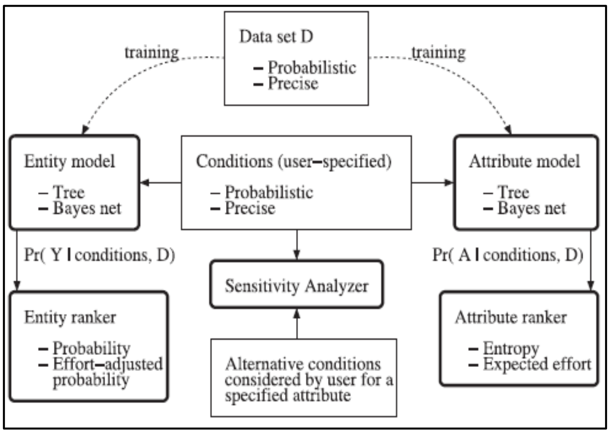


Fig. 4: FBR System Overview

B. *System Framework*

FBR System makes several contribution which aimed at improving basis for scientific search in databases. Scientific search normally include uncertainty in data and query also. FBR's another duty help the user to check the impact of giving a input, about which user is not sure. In some situation the impact of changing the input condition is high on result so it is correct to not change the conditions. To give risk calculation facility FBR System provide Sensitivity analysis.FBR System allow user to do operations on query output.

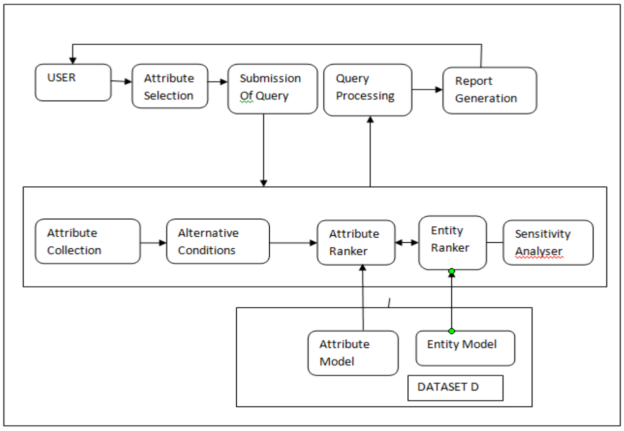


Fig. 5: System Architecture of FBR

FBR System response to the feature based input specified in terms of probability by calculating its rank according to entities and attributes. It will first trains the models of entity ranker and attribute ranker using training dataset. Then it become able to calculate the respective rankings. Sensitivity analysis can be done only if user demands. It suggest new conditions that will useful to improve the quality of result. It computes the sensitivity of result. If user modify the input conditions then FBR System provide the sensitivity of output, so user can avoid the modification in to the input if it have high impact on result. In fig 3 the architecture of FBR System is shown. The user give an input in terms of attributes. User can select attributes. Then the FBR System process the attributes and turns the user input in to the query. The query will processed by FBR System .And the report will generated by FBR System. The Attributes and Enteritis will be ranked by the Attribute and Entity model respectively.

One of the important feature provided by FBR System is that it will process the user input which contain uncertainity.FBR System's front end turns the user's uncertainty into the probability values.Fror prediction FBR System uses the bagged tree ensembles.These trees can handle any attribute type. Bagged trees having a capability that it can accommodate over fitting and robust against noise. These trees silted on to the attributes. For classification these trees will used by FBR System.

1) *Entity Ranking*

Entity Ranker uses entity model to rank the entities at the query time. Entity model will trained by training dataset which can be precise or probabilistic.

$$T_e rank=n_{y^0+u} \qquad (1)$$

Here $n_{y=}$Total no of nodes accessed in adaptive tree.
u= Constant independent of model size.

2) Attribute Ranking

Attribute Ranker uses Attributes at the query time. Attribute model will trained by training dataset which can be precise or probabilistic.

$$T_a\text{rank}=n_{a^{\theta}+u+(m-k)}\ s_2\ n_{y^{\theta}} \qquad (2)$$

Here,$n_{a=}$Total no of nodes accessed in adaptive tree for unspecified attributes.

3) Total Computation Time

Total computation time is the addition of time required to rank the entities and time required to rank the attributes.

4) Sensitivity Analysis

Sensitivity will be analysed only if user will demand.

III. Related work

Imprecise data design and methods, i. Highlighted a number of ongoing research challenges related to PDBs,and keeped referring to an information extraction (IE) scenario as a running application to manage uncertain and temporal facts obtainedfrom IE techniques directly inside a PDB setting[3]. Current approaches for answering queries with imprecise constraints require user-specific distance metrics and importance measures for attributes of interest – metrics that are hard to elicit from lay users are described[4].

The new method to calculate the coefficient of correlation.The minkowskies andaverage precision methods of calculation of the distance between two rankings are Discussed in detail.Takes approach toward a new rank correlation coefficient,AP correlation (ap), that is based on average precision and has a probabilistic interpretation.[2] User can express the uncertainty through probability values.The front end of FBR System is able to convert the user input into the probability values. Techniques to implement Merlin is given, methods related to calculation of sensitivity are given. Ranking distance calculation techniques are described[1]

All the techniques related to the data mining such as classification,clustering,data analysis,regression given in detail. To store the data, bagged trees used. So the concept of bagged trees, adaptive trees and implementation of those trees explained briefly.[5]

The attribute model and entity model will trained by training dataset. The machine learning technique is used to train the attribute model and entity model. The machine learning algorithms are given in detailed.[6]

IV. Conclusion

The proposed System is a completely automated approach for handling the uncertainty of user about query input. While doing big data analysis the database have to support many users finding information they are looking for.

The FBR System is proposed for allowing user to explicitly express their uncertainty through probabilities. FBR is a new way to scientifically search the large database.

Acknowledgment

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thankful to Dr. M. U. Karat, Head of Department, Computer Engineering., MET College of Engineering and Research Center, Nasik. We also thank the anonymous reviewers for their comments.

References

[1] T. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[2] E. Yilmaz, J. A. Aslam, and S. Robertson, “A new rank correlation coefficient for information retrieval,” in Proc. 31st Annu. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 587–594.

[3] D. Suciu, D. Olteanu, C. Re, and C. Koch, Probabilistic Databases. San Rafael, CA, USA: Morgan & Claypool, 2011

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011.

[5] B. Qarabaqi and M. Riedewald, “User-driven refinement of imprecise queries,” in Proc. IEEE 30th Int. Conf. Data Eng., 2014, pp. 916–927.

[6] B. Qarabaqi and M. Riedewald, “Merlin:Exploratory analysis with imprecise queries”in IEEE Trans.Knowl. Data Eng., vol. 20, no. 2, pp.342 –355, Feb. 2016.