iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** X **Month of publication:** October 2017

DOI: http://doi.org/10.22214/ijraset.2017.10158

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X, October 2017- Available at www.ijraset.com

Preliminary Interpretation Using Ambiguous Queries

Shivani Surana[1], Nikita Jadhav[2], Ishali Warungase[3], Subodh Baheti[4], Sachin Ubarhande[5]
[1, 2, 3, 4] *B. E. Scholars Computer Engineering, Sandip Institute of Engineering & Management*
[5] *Assistant Professor, Computer Engineering, Sandip Institute of Engineering & Management*

Abstract - Preliminary interpretation supports exploratory search in large databases.
The user interacts with it by specifying probability distributions over attributes, which then expresses imprecise conditions about the entities of interest. Preliminary interpretation helps the user on the right query conditions by addressing three key challenges:

1) ***Efficiently computing results for an imprecise query.***
2) ***Gives out the result of the sensitivity value of individual and combines Queries.***
3) ***Suggesting ideas for respective attributes for the user.***

Keywords: Interactive data exploration and discovery, Probability Query, Imprecise Queries, Sensitivity Analysis.

I. INTRODUCTION

The main notion of preliminary interpretation using ambiguous queries is helping the user with imprecise queries. The existing system is not able to handle the uncertainty of ambiguous queries, whereas proposed system allows user to express the uncertainty through the probabilistic value.

We propose this project, a new approach for exploratory searching large databases. This project provides a novel method to interactively compose imprecise database queries with probabilistic conditions, while providing constant feedback to the user about the most likely results and the potential benefit and risk of each condition. This method is designed to accommodate uncertainty and imprecision in user-provided query conditions through two major technical contributions:

A novel notion of sensitivity to quantify the impact of uncertainty on the query result.

Fast algorithms for calibrated probability estimation that can adapt to a user-specified real-time constraint on system response time.

To illustrate the need for imprecise queries with probabilistic conditions, consider the following example motivated by collaboration with the Cornell Lab of Ornithology. Through hugely successful citizen science projects such as the Lab has collected more than 100 million reports of bird sightings, adding tens of millions annually. It wants to leverage this resource to help less experienced birders identify the species of a bird they observed. Assume each observation in the database specifies properties of the bird (e.g., species, size, color) and the observation event (e.g., location, weather, habitat, and features).

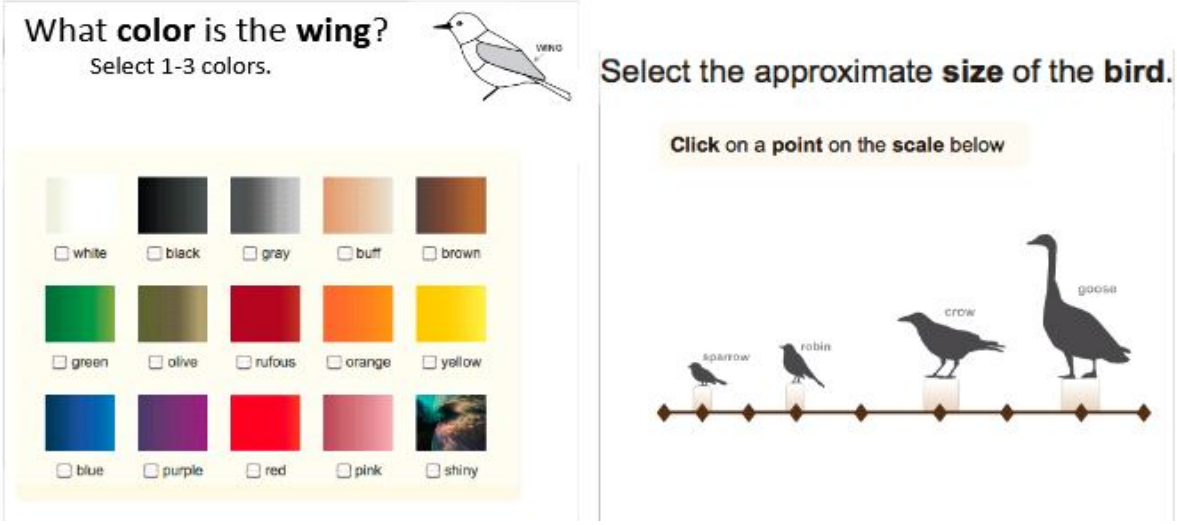


Figure (a): *Left* side image chooses the color and *Right* side image chooses the size

II. LITERATURE REVIEW

B. Qarabaqi and M. Riedewald, have proposed, "Client driven filtration of all the ambiguous queries" [1] The new method to calculate the coefficient of correlation.Traditional decision tree classifiers work with data whose values are known and precise. We then have to extend such classifiers to handle data with uncertain information, which originates from measurement/quantizationerrors, data staleness, and multiple repeated measurements.

S. Agarwal, B. Mozafari, A. Panda, H. Milner, S.T Madden and I.G.Stoica, have proposed, "Queries with bounded errors and bounded response times on big data" [2] theminkowskies and average 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), which is based on average precision and has a probabilistic interpretation.

D. Susie, D. Olteanu, C. Re, and C. Koch, have proposed, "Probabilistic Databases"[3] Imprecise data design and methods, Highlighted a number of ongoing research challenges related to PDBs, and keep referring to an information extraction (IE) scenario as a running application to manage uncertain and temporal facts obtained from IE techniques directly inside a PDB setting.

A. Abouzied, D. Angluin, M.QHellerstein, and A. Silbersc, have proposed, "Analyzing and verifying quantified Boolean queries by instances," [4] This project allows the client toanalyze and verifyall the quantified queries,a set of database queries known to be very difficult for all but the most expert users one can question the user on whether certain data objects are answers or not to her intended query. In this paper, we do the analysis of the number of questions needed to learn or verify the horn queries, a special class of Boolean quantified queries whose underlying form is conjunctions of the particular quantified by the Horn expressions. We also provide an optimal polynomial-question and tractable time learning and verification algorithms for two subclasses of the class qhorn with upper constant limits on a query's causal density.

I. F. Cruz, C. Stroe, and M.J.Palmonari,have proposed, "Interactive user feedback in ontology matching using signature vectors,"[5]When compared to a gold standard of the set of mappings that are generated using an automatic ontology matching process which is neither complete nor are the individual mappings always correct. Particularly as given the explosion in thesize, number and lastly the complexity of available ontologies, the domain experts no longer have the capability to create ontology mappings without considerable effort. Hence, we then present a solution to this problem that consists of making the ontology matching process interactive so as to incorporate user feedback in the loop. We approach clusters mappings to identify where user feedback will be most beneficial in reducing the number of user interactions and system iterations. This feedback process has been already implemented in the Agreement Maker system and is supported by visual analytic techniques that help users to better understand the matching process. An experimental result using the OAEI benchmarks shows the effectiveness of this approach. We will depict how users can interact with this ontology matching process through the Agreement MakerUser interface to match real-world ontologies.

S.Tsang, B. Kao, K. Yip, W.-S. Hoy, and S. D. Lee,have proposed ," Decision trees for uncertain data" [6]- An Automatic Interactive Data Exploration framework that iteratively steers the user towards interesting data areas and predicts a query that retrieves his objects of interest.

E. Yilmaz, J. A. Aslam, and S.T.PeterRobertson,have proposed," A new rank correlation coefficient for information retrieval"[7]- 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. We present AIMQ, a domain and user independent approach for answering imprecise queries over autonomous Web databases.

A. Paameswaran, A. D. Sarma, H. Garcia-Molinaand J.Q.Widom, have proposed, "The people who have assisted the graph search: It's okay to ask questions,"[8] As per thethe DAG i.e. directed acyclic graph(tree) with some target node(s), we consider the problem of finding the resultant target node(s) by asking an omniscient human questions of the form "Is there a target node that is reachable from the current node or not?". In this vivid problem,we have applications in many domains that can utilize human intelligence, including curation of all the possible hierarchies, debugging workflows, image segmentation and media refinement and categorization, interactive findings and filter synthesis. To our concern of knowledge, this work provides the first formal algorithmic study for the optimization of human computation for this problem. We also do the comparison of the performance of our algorithm against other algorithms, for the problem of webpage segmentation and categorization on the basis of real taxonomy. The provided framework and algorithms can be used in the design of an optimized for crowd-sourcing platforms such as Mechanical Turk.

K. Chen, H. Chen, N. Conway, M. Hellerstein, and T. S. Parikh, have proposed, "The User-driven Improving data quality with dynamic forms," [9]The dynamic form of theQuery is one of the most widely used user interfaces for querying databases. These kinds of traditional query forms are designed and pre-defined by developers or DBA in various information management systems. By this rapid development of web segmentation andmodern databases become very large and complex. Dynamic queries are a novel

approach to information seeking that may enable users to cope with information overload. They also allow users to visualizean overview of the database, conveniently filter out unwanted information. This paper thus results in proposing the Dynamic Query form, whic sacurious database query form interface, which is also able to dynamically generate query forms.

S. Branson Setal, have proposed, "The Visual identification with living creatures in the loop," [10]The user has a deep sense and understanding of the task, but limited memory and speed. The system has limited context and understanding, but powerful algorithmic resources. As a result, where the system cannot succeed reliably, the user steps in to guide the process which is not automatic, but manual.This tradeoff between the system and the user both defines and limits human-computer interaction. If the system cannot reliably support a task or goal, and if the user is unwilling to do it manually, the user may abandon the system. As a result, most interactive systems only attempt tasks they can safely automate. For example, consider the word processor — likely one of the most heavily-used and heavily-designed interactive systems of all time.

III. MATHEMATICAL MODEL

The complete system S can be represented in context of input given, functions performed and outputs generated.

S = {I,O,F}

where,

I : Input : {i1,i2..iN}

Where,

i1= 1st attribute.

i2= 2nd attribute.

O:Output: {Q, S, QP, SS, C, T}

Where,

Q=Query Generation from user input (Attribute collection)

S=Sensitivity value of specified attribute.

C=cost

T=Interactive response time to query response.

F:Functions: {Q, QP, SS, Re, QT}

Q=Query generation from user input

QP=Find Probabilistic Data on Uncertain Data.

SS=Stretching Shrink technique finds relaxations and contractions based on user feedback.

Re=Risk estimation, perform sensitivity quantifies the risk of a condition.

QT=Perform classification on Query Time.

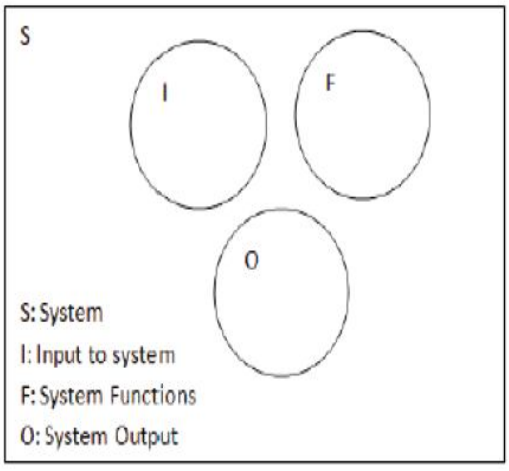


Figure 1: Vein Diagram

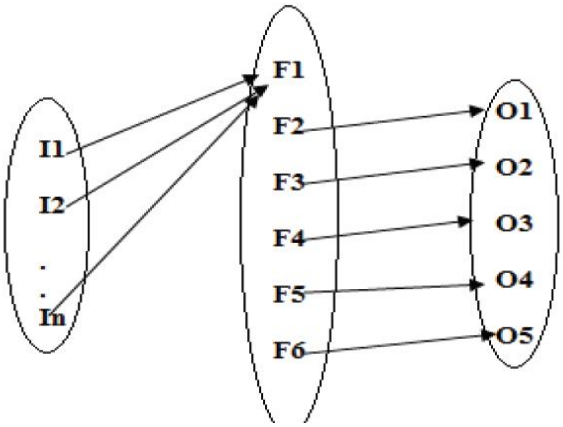


Figure 2: Functional Dependencies

IV. CONCLUSION AND FUTURE WORK

We proposed the novel notion of sensitivity to help user understand the potential risk of specifying a condition that she is not certain about. Knowing how sensitive the resulting ranked list is to the value of the condition, the user can decide if she wants to modify or eliminate the condition. We proved an important monotonicity property and proposed an algorithm that exploits it for more efficient estimation of sensitivity by "sampling from the edge". Our results also indicate that it is unlikely for a more efficient general sensitivity computation algorithm to exist. Using experimental results we showed how our efficient algorithm works in practice.

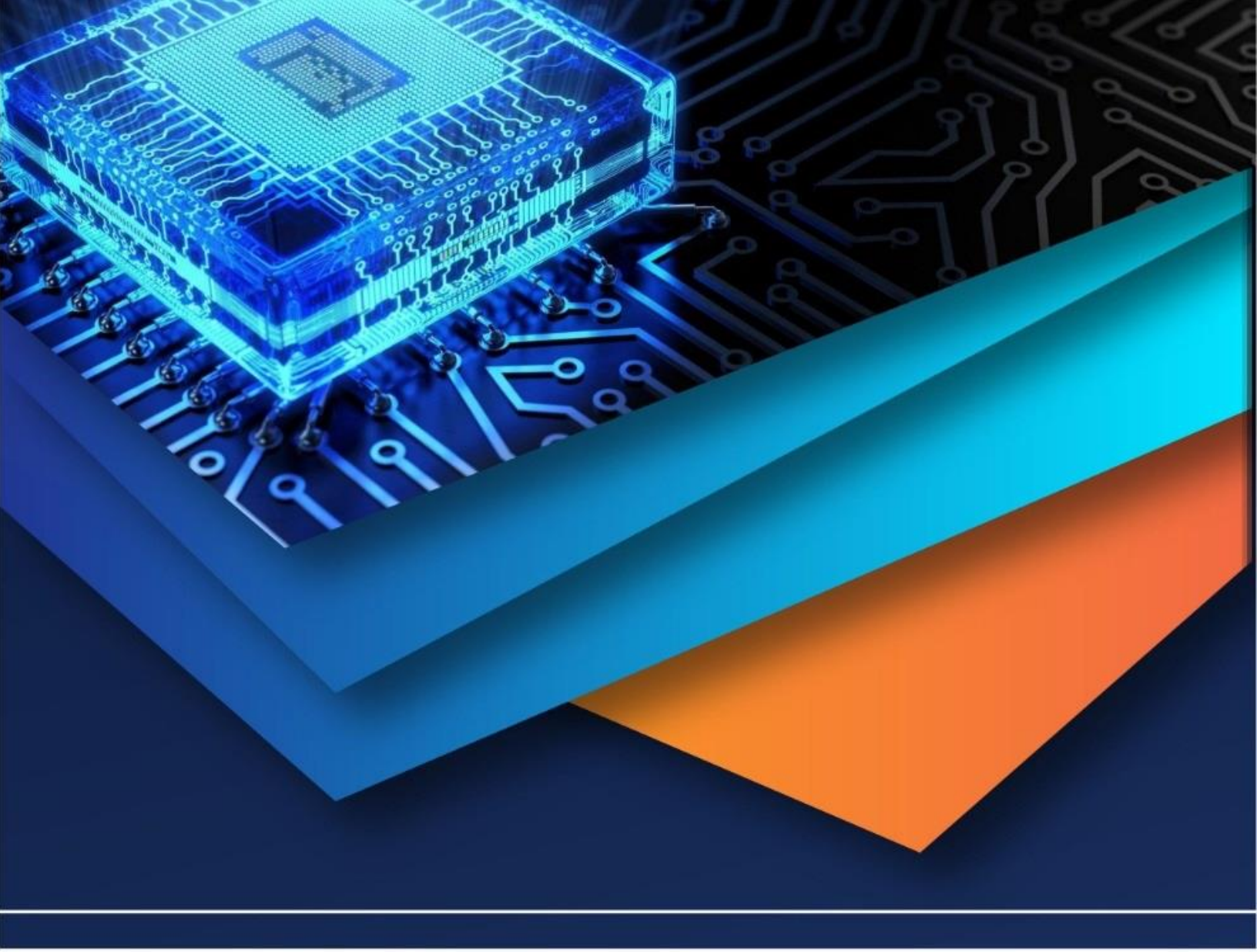
We conducted experiments to show that reduced-feature models are more calibrated than the complete models. An analytical model to explore this finding is a possible future work in this field. To gain wider acceptance for big data analysis and exploratory search, databases have to support a broad spectrum of users in finding the information they are looking for. We proposed Merlin as a novel framework to allow users to interactively compose imprecise queries by explicitly expressing the uncertainty through probabilistic conditions. We introduced the desired

Functionalities of Merlin and provided analytical and experimental results to show how Merlin supports them. The probabilistic framework is the core to all functionalities supported by Merlin.

This framework is needed, because imprecision can be present both in the data and the user-provided query conditions. One of the functionalities supported by Merlin is to provide constant feedback to the user, presenting a ranked list of the entities based on their likelihood of being the entity of interest, given the specified conditions by the user. We showed how to rank the entities to minimize the expected user effort.

REFERENCES

[1] B. Qarabaqi and M. Riedewald, "Client drivenfiltration of all the ambiguous queries," IEEE 30th Int. Conf. Data Eng., pp. 916927, 2015.
[2] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S.T. Madden, and I.G.Stoica,"Queries which are bounded errors and bounded response times on very big data" ACM European Conf. Computer. Syst., pp.2942, 2014.
[3] 39, 2013
[4] in Proc. 32nd Symp. Conf. , pp. 49–60, 2013,
[5] I. F. Cruz, C. Stroe, and M.J.Palmonari, "Interactive user feedback in ontology matching using signature vectors," in Proc. IEEE 11thInt.Conf., pp. 1321–1324, 2012
[6] S.Tsang, B. Kao, K. Yip, W.-S. Hoy, and S. D. Lee, "Decision trees for uncertain data" IEEE Trans. Knowl. Data Eng., no. 1, pp. 6478, Jan.2011.
[7] E. Yilmaz, J. A. Aslam, and S.T.PeterRobertson, "A new rank correlation coefficient for information retrieval" 35th Annual ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 587594, 2011.
[8] A. Parameswaran, A. D. Sarma, H. Garcia-Molina, and J.Q.Widom, "The people who have assisted the graph search: its okay to ask questions," Proc. VLDB Endowment, vol. 4, no. 5, pp. 267–278, 2011.
[9] K. Chen, H. Chen, N. Conway, J. M. Hellerstein, and T. S. Parikh, "The User-driven Improving data quality with dynamic forms," IEEE Trans. Data Conf., vol. 30, pp. 1138–1154, Aug. 2011.
[10] S. Branson et al., "The Visual identification with living creatures in the loop," Proc. 11th Eur. Conf. Computer, pp. 438–451, 2010.
[11] EBird (http://ebird.org)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY