International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 08 | Nov-2015 www.irjet.net p-ISSN: 2395-0072

Improved Search Results Of Keyword Query Using Data Imputation Approach

Priya Pujari [1], Arti Waghmare[2]

[1] M . E student, Department of Computer Engg, Dr. D.Y. Patil School of Engineering & Technology,Pune,Maharashtra
[2] Assistant Professor, Department of Computer Engg, Dr. D.Y. Patil School of Engineering & Technology, Pune,Maharashtra

---

Abstract - ***Keyword queries over databases offer simple access to data, but this frequently suffer from low quality of ranking. It would be beneficial to categorize queries that are likely to have the low ranking quality to improve the user satisfaction. For example, the system may recommend to the user alternate queries for such hard queries. In this report, the characteristics of hard queries are analyzed and a novel framework to compute the degree of difficulty for a keyword query above a database are proposed by considering both the structure and the content of the database and the query results. Additionally, Imputation is the process of supplanting missing information with the substituted values. Numerous current modern and exploration data sets contain missing values. Issues connected with missing values are loss of productivity, entanglements in taking care of and investigating the information and inclination coming about because of differences between missing and complete data.. The most existing imputation methods to string attribute values are inferring-based approaches with low imputation recall by just inferring missing values from the complete a portion of the data set. Recently, some retrieving-based methods are proposed to retrieve missing values from outer resources such as the World Wide Web but fails to respond to large number of search queries. So the approach is built called interactive Retrieving-Inferring data imPutation approach (TRIP), which is the interaction between the inferring-based methods and the retrieving-based methods. It performs retrieving and inferring alternately in filling in missing attribute values in a data set.***

Key Words: *Keyword query, Structured Robustness, improved searching , Data Imputation, Interactive Retrieving-Inferring , Query Prediction*

1. INTRODUCTION

Keyword query interfaces (KQIs) for databases have involved much attention in the past years because of their flexibility and availability in searching and exploring the data[1],[2]. From the time when any entity in a data set that holds the query keywords is a probable answer, keyword queries normally have numerous possible responses. KQIs necessarily identify the data needs behind keyword queries and result ranking so that the expected answers look at the top of the list. The retrieved results are analyzed to measure the power of a query over a database in retrieving the desired results. Some queries exhibits low ranking quality. Queries with low ranking quality are termed as difficult or hard queries. Identifying difficult queries will help in improving the performance by formulating methods to overcome the complexity involved in resolving the query. It is possible to optimize the query during query processing. Developing alternative queries or reformulating the query helps to overcome the difficulty involved with queries. This work mainly focuses on efficient prediction of difficult queries[3] and query results.

Also data imputation approach is considered while retrieving data from database. Data incompleteness is a prevalent data quality difficult in all types of databases. The procedure of filling in missing attribute values is called as Data Imputation[4]. To this point, quite imputation techniques have been established for missing quantitive information which is both numeric records such as temperature, age, salary etc., or categorical records with a relatively small scope of values [5] Simply limited consideration has been paid to non-quantitive information [6], which is string information with a large possibility of values .Conversely, string information takes up a huge part of the missing records in various databases. Present imputation methodologies to non-quantitive string information can be approximately put into two categories: (1) inferring- based methodologies, and (2) retrieving-based methodologies. Specially, the inferring-based methodologies find

replacements or approximations for the missing ones from the whole portion of the data set. Yet, they characteristically fall short in filling in single missing attribute values which do not occur in the whole part of the data set [7]. The retrieving-based methods resort to exterior resources intended for help. Based on the evidence that the missing records might exist at some exterior data sources over the WWW (World Wide Web), particular work has been directed to yield missing values from web tables and web lists but fails to respond to large number of search queries. So here investigating the interaction between the inferring-based methods and the retrieving-based methods and propose an interactive Retrieving-Inferring data imputation approach (TRIP) [8], which performs retrieving and inferring alternately in filling in missing attribute values in a data set. TRIP identifies an optimal retrieving-inferring scheduling scheme in Deterministic Data Imputation (DDI) and also identifies an expected-optimal scheme in τ —SDI. Using this method one can achieve high recall, better efficiency and accurate query results.

1.1 The design objectives and goals

I. To effectively predicts the ranking quality of representative ranking algorithms.
II. To design search schemes which allow keyword query search and relevant data retrieval.
III. To effectively perform data imputation.

1.2 Motivation

Keyword query interfaces for databases have attracted more attention in the last few years due to their flexibility and availability in searching and discovering the data. Data incompleteness is also a prevalent data quality problem in all types of databases. The process of filling in missing attribute values is well-known as Data Imputation which is more important to get better results on keyword queries.

2. BACKGROUND

A. Keyword Query:
keyword queries normally have numerous possible responses. Databases comprise entities, and entities comprise attributes that yield attribute values. Certain difficulties of answering a query are as follows:
(1) Unlike queries in languages like MySQL, normally users do not specify the chosen schema element for every query term. For example, query Q1: Godfather on the IMDB record does not stipulate if the user is interested in movies whose name is Godfather or movies disseminated by the Godfather Corporation. Therefore, a KQI essentially find the preferred attributes related with every term in the query.
(2) The representation of the output is not quantified, i.e., users do not provide sufficient information to choose exactly their favorite entities.

It is important for a KQI to recognize such queries and warn the user or employ alternative techniques like query reformulation or query suggestions [9]. It may also use techniques such as query results diversification [10]

B. Ranking Technique:
Ranking of the query results gives us Top-K entities. Here use the ranking algorithm called Probabilistic Retrieval Model for Semistructured Data**(**PRMS) . It assigns each attribute a query keyword-specific weight, which specifies its contribution in the ranking score.[11]

C. Noise generation in database:
Noise generation model will show that every attribute value is corrupted by a grouping of three corruption stages: on the value itself, its attribute and its entity set. In this model, the focus simply on the noise introduced in the content (values) of the database.

D. SR Score:
Structured Robustness (SR) score, which measures the difficulty of a query based on the differences between the rankings of the same query over the original and noisy (corrupted) versions of the same database and Spearman rank correlation is used to compute the similarity of the answer lists of both database.

E. Interactive Retrieving & Inferring data imPutation approach(TRIP):
TRIP is an interactive retrieving and inferring data imputation approach, which can profit from the high recall of retrieving-based approach and the efficiency of inferring-based approach. While an inferring step in TRIP fills in all currently inferable missing values to the greatest extent, the succeeding retrieving step retrieves a set of selected missing values that make some unfilled missing values become inferable for the next inferring step. Inferring and retrieving are alternately performed until no more missing values can be imputed. The interaction between retrieving and inferring can be simply represented by a sequence of missing value sets. We call this sequence of missing value sets as scheduling scheme. TRIP able to identify optimal scheduling scheme in Deterministic Data Imputation (DDI) and expected-optimal scheme in τ —constrained Stochastic Data Imputation (τ-SDI)

3. LITERATURE SURVEY

A. Difficult Keyword Query over database

In paper [12] Venkatesh Ganti et al. suggest overall framework that can increase an existing search interface

by interpreting a keyword query to an arranged query. Precisely, they control the keyword to attribute value relations exposed in the results resumed by the novel search interface.

In paper [13], Nikos Sarkas et al. learn latent organized semantics in web queries and yield Structured Annotations for them. Authors deliberate an annotation as a mapping of a query to a table of arranged data and attributes of this table. Given a collection of structured tables, they existent a fast and scalable tagging techniques for obtaining all conceivable annotations of a query over these tables.

Kevyn Collins-Thompsonan and Paul N. Bennett [14] introduce novel models and illustrations for approximating two significant measures of query presentation: query difficulty and expansion risk. Their effort brings collected features from preceding studies on query effort based on deviations between language representations of the query, collection and initial outcomes.

Oren Kurland et al. [15] presented a fundamental probabilistic prediction structure. By using their framework, they develop and describe numerous previously projected prediction approaches that might appear totally different, but produce to share the similar formal basis. As well the framework is used to formulate novel prediction methodologies that leave behind the state-of-the-art.

In paper [16], Shiwen Cheng et al. investigate the characteristics of hard queries and suggest a new framework to calculate the degree of difficulty for a keyword query over a database, seeing both structure and content of the database and the query outcomes.

In paper [17], Arash Termehchy et al. introduced and describe independence of design, which captures property for Schema free query interfaces (SFQIs). Authors establish a theoretical structure to compute the design amount independence delivered by an SFQI. Authors illustrate that most existing SFQIs deliver a limited degree of design independence.

Probabilistic Retrieval Model for Semistructured Data (PRMS) services a language model methodology to search above structured information. It evaluates the language model of every attribute value leveled by the language model of its attribute. It gives every attribute a query keyword with specific weight, which states its involvement in the ranking scores [10].

B. Data Imputation

Paper [18] studies a novel problem, the interaction among record matching and data repairing. Wenfei Fan et al. suggests a uniform framework that effortlessly combines repairing and matching procedures, to cleared a database based on reliability constraints, master data and matching rules.

Ravi Gummadi et al. [19] presented a combined methodology that supports intelligent recovery above fragmented web databases by mining and by means of inter-table needs.

Many works has been conducted to harvest missing values from either web lists or web tables [20]. More recently, a general web-based retrieving approach was proposed to retrieve missing data from all kinds of web documents [21].

In paper [22], Huawen Liu et al. suggest anomaly elimination and learning algorithm below the framework of kNN. The main characteristic of technique is that the suggestion of eliminating anomalies and predicting class labels of unseen examples is adjacent nearest neighbors, somewhat than k-nearest neighbors.

In paper [23], Mohamed Yakout et al. describes the INFOGATHER system to mechanize information gathering responsibilities, like augmenting entities with attribute values and searching attributes, via web tables.

4. PROPOSED SYSTEM

Existing system is not responsive on database missing values. Existing system projected a framework that focus only on difficult keyword searching techniques over database. Thus if one can consider that there is missing data into database then efficiency of existing framework gets reduce since results may not get as expected.Here we analyzes the individualities of difficult queries over databases and suggests a novel technique to detect such queries. The main challenge in using the Ranking Robustness Principle for databases is to describe data corruption for structured records. Thus database using a reproductive probabilistic model based on its structure chunks, which are terms, attribute values, attributes, and sets of entities is modeled. Additionally examines the interaction among the inferring-based approaches and the retrieving-based approaches. Retrieving a lesser number of designated missing values can significantly improve the imputation recall of the inferring-based techniques. With this perception, an inTeractive Retrieving-Inferring data imPutation approach (TRIP) is proposed, which achieves retrieving and inferring interchangeably in filling in missing attribute values in a database. To guarantee the high recall at the least cost, TRIP faces a challenge of choosing the least number of missing values for retrieving to exploit the number of inferable values. The proposed resolution is capable to identify an ideal retrieving-inferring scheduling structure in Deterministic Data Imputation (DDI), and the optimality of the created system is theoretically examined with proofs.

4.1 Architectural Diagram:

The following Fig 1.shows overview of system architecture

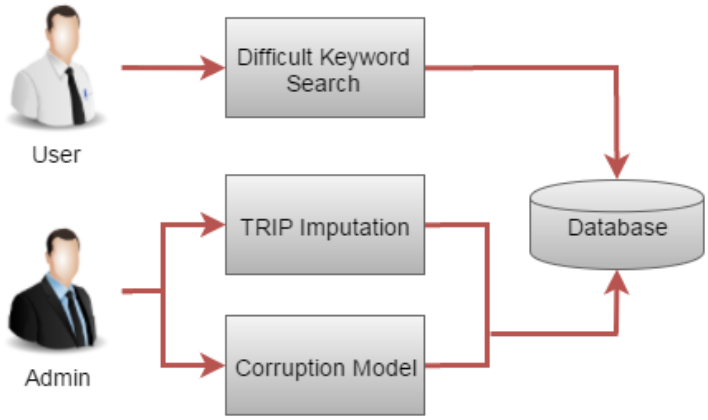

Fig 1. System architecture

Structured Robustness (SR) score, which measures the difficulty of a query based on the differences between the rankings of the same query over the original and noisy (corrupted) versions of the same database and TRIP performs retrieving and inferring alternately in filling in missing attribute values in a dataset.By considering both data imputation TRIP method and prediction of difficult keyword query using SR algorithm we can get better results and high efficiency.

5. CONCLUSION AND FUTURE WORK

Based on framework, a novel algorithm is proposed that efficiently predict the effectiveness of a keyword query over a database, using the ranking robustness principle. Additionally, propose a hybrid retrieving-inferring data imputation approach TRIP to alternately perform retrieving and inferring in imputing missing values in a database.

Future work may use inference rules and retrieving queries corresponding to other attribute dependencies, and apply TRIP to some other SDI scenarios. Also we can use another algorithm which performs better than SR algorithm.

6. ACKNOWLEDGEMENT

The work described in this paper is supported by the Department of Computer Engineering of D.Y.Patil School of Engineering and Technology, Pune. The authors would like to thank all staff members of Computer Engineering department for their valuable support.

7. REFERENCES

[1] V. Hristidis, L. Gravano, and Y. Papakonstantinou, "Efficient IRstyle keyword search over relational databases," in *Proc. 29th VLDB Conf.*, Berlin, Germany, 2003, pp. 850–861.

[2] Y. Luo, X. Lin, W. Wang, and X. Zhou, "SPARK: Top-k keyword query in relational databases," in *Proc. 2007 ACM SIGMOD*, Beijing, China, pp. 115–126.

[3] Shiwen Cheng, Arash Termehchy, and Vagelis Hristidis, "Efficient Prediction of Difficult Keyword Queries over Databases", vol. 26, no. 6, JUNE 2014.

[4] G. Batista and M. Monard. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6):519–533, 2003

[5] J. Tian, B. Yu, D. Yu, and S. Ma. Missing data analysis: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering. Applied Intelligence, pages 1–13, 2013.

[6] Q. Wang and J. Rao. Empirical likelihood-based inference under imputation for missing response data. The Annals of Statistics, 30(3):896–924, 2002.

[7] Z. Li, M. A. Sharaf, L. Sitbon, S. Sadiq, M. Indulska, and X. Zhou. Webput: Efficient web-based data imputation. In WISE, pages 243–256, 2012.

[8] Zhixu Li, Lu Qin, Hong Cheng, Xiangliang Zhang, and Xiaofang Zhou, "TRIP: An Interactive Retrieving-Inferring Data Imputation Approach," IEEE Transaction 2015.

[9] A. Nandi and H. V. Jagadish, "Assisted querying using instantresponse interfaces," in *Proc. SIGMOD 07*, Beijing, China, pp. 1156–1158.

[10] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, "DivQ: Diversification for keyword search over structured databases," in *Proc. SIGIR' 10*, Geneva, Switzerland, pp. 331–338.

[11] J. Kim, X. Xue, and B. Croft, "A probabilistic retrieval model for semistructured data," in *Proc. ECIR*, Tolouse, France, 2009,
pp. 228–239.

[12] V. Ganti, Y. He, and D. Xin, "Keyword++: A framework to improve keyword search over entity databases," in Proc. VLDB Endowment, Singapore, vol. 3, no. 1–2, pp. 711–722, Sept. 2010.

[13] N. Sarkas, S. Paparizos, and P. Tsaparas, "Structured annotations of web queries," in Proc. ACM SIGMOD Int. Conf. Manage. Data, Indianapolis, IN, USA, pp. 771–782, 2010.

[14] K. Collins-Thompson and P. N. Bennett, "Predicting query performance via classification," in Proc. 32nd ECIR, Milton Keynes, U.K., 2010, pp. 140–152.

[15] O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, and O. Rom, "Back to the roots: A probabilistic framework for query performance prediction," in Proc. 21st Int. CIKM, Maui, HI, USA, 2012, pp. 823–832.

[16] S. Cheng, A. Termehchy, and V. Hristidis, "Predicting the effectiveness of keyword queries on databases," in Proc. 21st ACM Int. CIKM, Maui, HI, 2012, pp. 1213-1222.

[17] A. Termehchy, M. Winslett, and Y. Chodpathumwan, "How schema independent are schema free query

interfaces?" in Proc. IEEE 27th ICDE, Hannover, Germany, 2011, pp. 649–660.
[18] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu. Interaction between record matching and data repairing. In SIGMOD, pages 469–480, 2011.
[19] R. Gummadi, A. Khulbe, A. Kalavagattu, S. Salvi, and S. Kambhampati. Smartint: using mined attribute dependencies to integrate fragmented web databases. Journal of Intelligent Information Systems, pages 1–25, 2012.
[20] Z. Li, M. A. Sharaf, L. Sitbon, S. Sadiq, M. Indulska, and X. Zhou. Webput: Efficient web-based data imputation. In WISE, pages 243–256, 2012.
[21] Z. Li, M. A. Sharaf, L. Sitbon, S. Sadiq, M. Indulska, and X. Zhou. A webbased approach to data imputation. WWW, 2013
[22] H. Liu and S. Zhang. Noisy data elimination using mutual k-nearest neighbor for classification mining. Journal of Systems and Software, 85(5):1067–1074, 2012.
[23] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In SIGMOD, pages 97–108, 2012.