International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Survey Paper on Retrieval of Deterministic Data with Ranking Strategy

Ashish S Mutrak[1], Kishor Shedge[2]

[1]Student, Master of Engineering, Department of Computer Engineering, Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

[2]Assistant Professor, Department of Computer Engineering, Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

Abstract: ***Now a days because of the widespread use of Internet users prefers to store the data on web applications such as cloud. Most of the storage applications stores the deterministic data in legacy systems. This paper focuses on the problem of deterministic data & enables to be stored in legacy systems. The probabilistic data may be generated by automated data analysis techniques. Our goal is to generate the deterministic representation of probabilistic data that optimizes the quality of the end- application. The determination of probabilistic data can be done by using query aware strategy and ranking algorithms. These strategies helps to generate more precise representation of the data as the selection of data is done by triggering a query on specific items and also an index ranking is applied to the query selected attributes for more accurate representation. We have also highlighted the advantages of using these techniques over the traditional approaches.***

Keyword: Probabilistic data, legacy systems, triggers, deterministic data

1. Introduction

With rapid & widespread acceptance of internet and cloud technology user often prefers to store their data on different web applications. Mostly the user data is generated automatically through variety of online means such as signal processing techniques, data analysis and mining methods before it is stored on web applications.

For example the Exif data which can automatically stores the geographical attributes of an image. If a user directly uploads all this information on a storage system, it may create duplicated entries. This duplication needs to be determinized before forcing it on the storage or database. There are many traditional approaches used to design these kind of systems where a Top-1 and a All technique is used where it selects a value which is more relevant to the query. But many times these techniques leads to less than a highest quality output.

Our objective is to design a customized approach that selects a more specific representation and in turn will optimize the quality of end application. The use of triggers in our paper chooses the best deterministic representation.

Unlike determinizing an answer to a query, our goal is to consider the effective cost of finding an item in a dataset. The principle of ranking is used to find the most probable data from a dataset where a query generates more than required relevant results. The frequent item set is then given to branch bound algorithm, which finds the most probable solution to the query. Many web pages which stores the contents generated with live action are dynamic in nature, meaning that they gradually updates the information. For this dynamic updates huge datasets are used.

Internet has a wide source of knowledge stored which can be used by the user or administrator to maintain the huge databases. This is done with the cloud storage systems. Following are few ways from which an optimal solution can be obtained:

1) The searching method can be improved by maintaining the frequent search parameters.
2) Knowledge extraction techniques can be imposed.

Knowledge based searching should be done. Let us consider the example two users were searching for a specific data virus affected system, one of the user need the data that should be related to medical terms and another user need the data that should be related to computer system but both the user express their query in the same manner. For some queries, everyone who issues the query is looking for the same thing. For many other queries, different people want very different outcomes even though they express their need in the same manner. By studying the browsing history and frequent visited pages the user can analyze the exact need.

1.2 Justification of the Problem

Two basic approaches for the determinization problem are Top-1 and All technique are used. In Top-1 approach the most probable value is selected and in All technique all the possible values with non-zero probability are selected.

For example an exact number finder system which generates a single value can be viewed using a top-1 method. For a Heartbeat scanner which must include all the possible values, All technique could be useful for that case. However, such approaches being agnostic to the end-application often lead to suboptimal results. A better approach is to design customized determinization strategies that select a determinized representation which optimizes the quality of the end-application.

1.3 Objectives of Proposed system

This system focuses on producing a deterministic depiction that optimizes the quality of answers to queries. The system also aims to select a determinized representation which

optimizes the quality of the end-application, also the workload is fired with the query it finds the best deterministic representation that reaches the near optimal quality.

2. Related Works

1. In the last decade, a lot of work has been done in the field of similarity query processing with the focus on management and processing of uncertain data. Thereby,the development of eficient and effective approaches providing probabilistic query results were of main interest. A survey of the research area concerning uncertainty and incomplete information in databases is given in [12] and [18].

Recently a lot of work has been published in the area of management and query processing of uncertain data in sensor data-bases [14] and especially in moving object environments [15], [19]. Similar to the approach presented in this paper, the approaches in [13],[14],[15],[19] model uncertain data by means of probabilistic density functions (pdf). In [19], for instance, moving objects send their new positions to the server, if their new positions considerably vary from their last sent positions. Thus, the server always knows that an object can only be a certain threshold value away from the last sent position. The server, then, as signs a pdf to each object reecting the likelihood of the objects possible positions. Based on this information the server performs probabilistic range queries.

2. Likewise, in [15] an approach is presented for probabilistic nearest neighbor queries. Note that both approaches assume non-uncertain query objects, and thus, they cannot be used for queries where both query and database objects are uncertain. Queries that support uncertain database objects as well as uncertain query objects are very important as they build a foundation for probabilistic join procedures. Most recently, in [17] a probabilistic distance range join on uncertain objects was proposed. Instead of applying their join computations directly on the pdfs describing the uncertain objects, they used sample points as uncertain object descriptions for the computation of the probabilistic join results.

Furthermore, most recently [16] an approach was proposed dealing with spatial query processing not on positionally uncertain data but on existentially uncertain data. This kind of data naturally occurs, if, for instance, objects are extracted from uncertain satellite images. The approach presented in this paper does not deal with existentially uncertain data but with positionally uncertain data which can be modelled by probability density functions or are already given as probabilistic set of discrete object positions similar to the approach presented in [17].

3. Dmitri V. Kalashnikov addresses the poor quality of annotations by incorporating outside semantic knowledge to improve interpretation of the recognizers output, as opposed to blindly believing what the recognizer suggests. Most speech recognizers provide alternate hypotheses for each speech utterance of a word, known as the N-best list for the utterance. They exploit this fact to improve interpretation of speech output. The goal is to use semantic knowledge in traversing the search space that results from these multiple alternatives in a more informative way, such that the right annotation is chosen from the N-best list for each given utterance. We show that by doing so, we can improve the quality of speech recognition and thereby improve the quality of the image tag assignment. 3.Jia Li James Z. Wang uses pictorial information of each image is summarized by a collection of feature vectors extracted at multiple resolutions and spatially arranged on a pyramid grid. The 2D MHMM fitted to each image category plays the role of extracting representative information about the category. In particular, a 2D MHMM summarizes two types of information: clusters of feature vectors at multiple resolutions and the spatial relation between the clusters, both across and within resolutions. As a 2D MHMM is estimated separately for each category, a new category of images added to the database can be profiled without repeating computation involved with learning from the existing categories.

4.Changhu Wang,Feng Jing, Lei Zhang Hong-Jiang Zhang used, a set of candidate annotations for the query image need to be identified. We deal with Web images and non-Web images in different ways. Second, the RWR algorithm is used to re-rank the candidate annotations. Finally, the top ranked annotations will be reserved as the final annotations.

5.Rabia Nuray-Turan, Dmitri V. Kalashnikov, Sharad Mehrotra, And Yaming Yu developed a system to e_ciently compute answer to selection queries that would maximize quality where we fix the quality metric to be the F-measure. One of the challenges is that the ground-truth answer is unknown to the algorithm beforehand, so that it cannot measure quality of different answers directly. The idea of our solution, to which we refer as Maximization of Expectation (MoE), is to find answer optimal solution that would maximize the quality in the expected sense. 6.Sumit Bhatia, Debapriyo Majumdar, Prasenjit Mitra proposed a system to offer query suggestions to the user even in scenarios where query logs are not available.

In the absence of query logs, we adopt a document-centric approach by utilizing the documents in the corpus itself to generate query suggestions on the y. We extract and index phrases from the document corpus and when the user starts typing a query, we utilize these phrases to complete the partial user query. The completed queries are then offered as suggestions to the user.

3. Proposed System

3.1 Architecture of the Proposed System

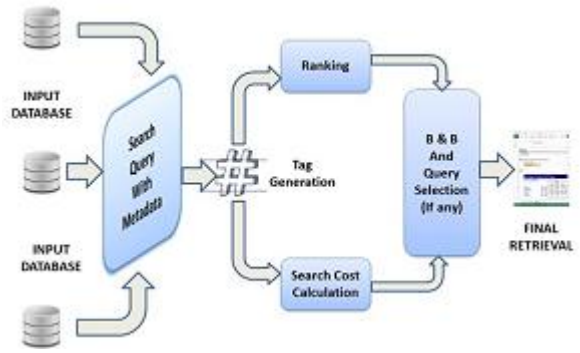

Figure 3.1: Architecture of System

The system accepts the data in either an online or an offline way. The datasets may correspond to a system which works only on deterministic datasets. The search query is then fired on the items which generates a set of relative items. These relevant items may get duplicated. This nature is purely dependent on the query fired on the database. The selected items may mark as a false positive and false negative entities. The estimated cost of finding an entity is calculated. The relevant items are then ranked depending upon their probability of searching. And after a Branch and Bound method is applied on the ranked entities which leads to find most appropriate determination of objects.

For representation of each object we can define quality metric of the selected dataset values with regard to a given query workload. The choice of the metric depends upon the end application.

The object should be associated with the following information:

1)The set of tags provided by some data processing technique, e.g., speech recognition, or entity resolution.
2)Probabilities provided by the data processing techniques.

Object	CAR	ROAD	BUILDING
Probability			

Figure 3.2: Object Representation

Some application which supports Triggers / Selection Queries the cost can be calculated as False Positive False Negative cost.

False Positive: Objects that satisfies the query but does not belongs to the Ground truth. In this case it will retrieve the irrelevant objects.

False Negative: Objects that does not satisfies the query but belongs to the Ground truth. In this case it may miss the relevant objects.

As we are applying the threshold values we cannot compute the cost directly as the threshold values may vary for different datasets. And also our main aim is to minimize the cost. The idea is to select the objects based on their cost. This cost calculation is only for the uncertain objects. Along with the calculation of cost each retrieved object needs to be ranked as the user is having a provision to select the objects based on their cost or based on their ranking. This will make the retrieval process much easier to understand to implement also. The ranking to the objects is done on the basis of their probabilities.

4. Conclusion

In this paper we have considered the problem of retrieval of deterministic data. Our goal is to generate a deterministic representation that optimize the quality of answers to queries or triggers that execute over the deterministic data representation and to store this data on legacy systems which accepts only deterministic input. We have proposed a solution that optimizes the result using ranking in object retrieval. This technique is much more faster than enumeration based optimal solutions and also achieves almost the same quality as that of the optimal solution.

5. Acknowledgement

I would like to express my profound gratitude and deep regard to my Project Guide Prof. K.N.Shedge, for his exemplary counsel, valuable feedback and constant fillip throughout the duration of the project. His suggestions were of immense help throughout my project work. Working under him was an extremely knowledgeable experience for me.

References

[1] D. V. Kalashnikov, S. Mehrotra, J. Xu, and N. Venkatasubramanian, A semantics based approach for speech annotation of images, IEEE Trans. Knowl. Data Eng., vol. 23, no. 9, pp. 13731387, Sept. 2011.
[2] . Li and J. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. Pattern Anal. Mach.Intell., vol. 25. 10751088, Sept. 2003.
[3] C. Wangand, F. Jing, L. Zhang, and H. Zhang, Image annotation refinement using random walk with restarts, in Proc. 14th Annu.ACM Int. Conf. Multimedia, New York, NY, USA, 2006.
[4] J. Li and A. Deshpande, Consensus answers for queries over probabilistic databases, in Proc. 28th ACM SIGMOD-SIGACTSIGART Symp. PODS, New York, NY, USA, 2009.
[5] S. Bhatia, D. Majumdar, and P. Mitra, Query suggestions in the absence of querylogs, in Proc. 34th Int. ACM SIGIR, Beijing,China, 2011.
[6] I. Bordino, C. Castillo, D. Donato, and A. Gionis, Query similarity by projecting the query flow graph, in Proc. 33rd Int. ACMSIGIR, Geneva, Switzerland, 2010.
[7] A. Anagnostopoulos, L. Becchetti, C. Cas tillo, and A. Gionis, An optimization framework for query recommendation, in Proc. 3rd ACM Int. Conf. WSDM, New York, NY, USA, 2010.

[8] A. Rae, B. Sigurbjrnsson, and R. V. Zwol, Improving tag recommendation using social networks, in Proc. RIAO, Paris,France, 2010.
[9] B. Sigurbjrnsson and R. V. Zwol, Flickr tag recommendation based on collective knowledge, in Proc. 17th Int. Conf. WWW,New York, NY, USA, 2008.
[10] Abiteboul S., Hull R., Vianu V.: Foundations of Databases.Addison Wesley, 1995.
[11] Bhm, C., Pryakhin A., Schubert M.: The Gaus-Tree: Effcient Object Identification of Probabilistic Feature Vectors. ICDE06.
[12] Cheng R., Kalashnikov D.V., Prabhakar S.: Evaluating probabilistic queries over imprecise data. SIGMOD03.
[13] Cheng R., Kalashnikov D. V., Prabhakar S.: Querying imprecise data in moving object environments.IEEE Transactions on Knowledge and Data Engineering, 2004.
[14] Dai X., Yiu M., Mamoulis N., Tao Y., Vaitis M.: Probabilistic Spatial Queries on Existentially Uncertain Data.SSTD05.
[15] Kriegel H.-P., Kunath P., Pfeie M., Renz M.: Probabilistic Similarity Join on Uncertain Data. DASFAA06.
[16] Motro A.: Management of Uncertainty in Database Systems.In Modern Database Systems,Won Kim (Ed.), Addison Wesley, 1995.
[17] Wolfson O., Sistla A. P., Chamberlain S., Yesha Y.: Updating and Querying Databases that Track Mobile Units.Distributed and Parallel Databases, 7(3), 1999. W.,Chellappa R., Phillips P.J., Rosenfeld A.: Face Recognition: A literature survey.ACM Computational Survey, 35(4), 2000.

Author Profile

Ashish S Mutrak received the B.E. degree in Computer Engineering from K.K.Wagh Institute of Engineering Education and Research Center in 2011. He is currently pursuing Master's Degree in Computer.

Prof. K. N. Shedge is working as Assistant Professor in Sir Visvesvaraya Institute of Technology.