Survey on Query Aware Strategy for Determining Uncertain Probabilistic Data

Umesh Gorela[1], Bidita Hazarika[2], Abhinesh Tiwari[3], Priti Mithari[4]

Department of Computer Engineering, Savitribai Phule University of Pune , India[1,2,3,4]

***Abstract*— In this paper the main problem which is considered is the determinization of probablistic data which capable to enable such data which is to be stored in legacy systems that accepts only the deterministic input. Probabilistic data is generated by automated data analysis/enrichment techniques like entity resolution, information extraction, and speech processing. Legacy system which is used is corresponding to the pre-existing web applications like Picasa, Flickr etc. Our intention and the very goal is to generate a deterministic representation of probablistic data which optimizes the quality of the end-application built on determisnistic data. Exploring such a problem in the context of two very different data processing tasks-which can be also termed as triggers and selection queries. There by showing the approaches like threshholding or top-1 selection which is traditionally used for determinizing leading to suboptimal performance for such kind of applications .Instead developing a query-aware strategy and showing its various advantages over the existing solutions through a wide-ranging empirical evaluation over the real and synthetic datasets.**

***Index Terms*—Determinization, data quality, query workload, uncertain data.**

I. INTRODUCTION

With the introduction of cloud computing and the rapid increase of the use of web-based applications, people often save their data in many various existing web applications. Often, data of user is generated automatically through a variety of signal processing,query analysis /enrichment techniques before being stored in the various web applications. For example modern DSLR cameras support analysis of vision in order to generate tags such as indoors/outdoors, various scenery, landscape / portrait etc. Many modern photo cameras often have microphones for users to speak out a descriptive sentence which is then recognized by a speech recognizer to generate a set of tags to be associated with the picture [2]. The picture (along with the set of tags) can be seen in real-time using wireless connectivity to Web applications such as Flickr. Putting such data into web applications poses a challenge since such automatically generated content is often uncertain and may result in objects with probabilistic attributes. For example, vision analysis may result in tags with probabilities [3], [4], and, similarly automatic speech recognizer (ASR) may produce an N-best list or a confusion network of utterances [2], [3].This kind of probabilistic query must be "determinized" before being saved in legacy web applications. We refer to the difficulty of mapping probabilistic data into the similar deterministic representation as the *determinization* problem. Many such approaches for the *determinization* problem can be made. Two main approaches are the Top-1 and All techniques, where we choose the most probabilistic value / all the possible values of the attribute with the probability non-zero, respectively. For example, a speech recognition system that generates a single answer/tag for each expression can be seen as using a top-1 strategy. Another technique might be to choose a threshold τ and include each and every attribute values with a probability greater than τ . However, such approaches being doubted to the end-application often lead to suboptimal results. A better approach is to design custom determinization strategies that choose a determinized representation which optimizes the value of the end-application. Consider, for example, an end app that supports triggers/alerts on automated content generation. Examples of such an end-app includes publishing/subscribing system such as Google Alert, where people put their subscriptions in the form of index keywords (e.g.Gujarat earthquake) and predicts over a database (e.g. this data is video). Google Alert finds all corresponding data sets to the user based on the subscriptions. Now for example a video about Gujarat Earthquake is to be uploaded on YouTube. The video has a set of tags that were decided using either by automatically vision processing and/or by information retrieval techniques put over transcribed speech.

ISSN: 2278 – 7798

Such tools which may create tags with probabilities (e.g., “Gujarat": 0.8, “earthquake":0.4, “election": 0.6), while the important tags of the video could be “Gujarat" and “earthquake". The determinization procedure should link the video with suitable tags such that subscribers or the users who are really very much involved in the video (i.e., whose subscription includes the words “Gujarat Earthquake") are notified while others are not overwhelmed by immaterial data. Thus, in the given example , the determinization process should minimise metrics called as false positives and false negatives that result from a determinized representation of data. Now take a example of different application such as Flickr, to which pictures are uploaded automatically from modern cameras along with the tags that may be generated based on speech recognition or image enrichment techniques. Flickr supports effective retrieval based on photo tags. In such an application, people may have interest in selecting determinized representation that optimizes set-based quality metrics such as F-measure instead of minimizing false positives/negatives. In this paper, we study the difficulty of deteminizing datasets with probabilistic attributes (usually generated by automatically by data analyses/enrichment). Our approach exploits a workload of triggers/queries to choose the top deterministic representation for two types of applications– one that chains triggers on generated content and another that supports effective retrieval. Interestingly, the trouble of determinization has not been explored widely in the past. The most related research efforts are which explore how to give deterministic answers to a query (e.g. conjunctive selection query) over probabilistic database. Unlike the problem of determinizing an answer to a query, our aim is to determinize the data so as to enable it to be stored in legacy deterministic databases such that the determinized representation maximises the anticipated performance of queries in the future. Solutions in [5], [6] cannot be straightforwardly applied to such a determinization problem. Probabilistic data is studied in this paper; the works that are mostly related to ours is this project. They search how to determine answers to a query over a probabilistic data. In similarity, we have interest in best deterministic representation of data (and not Determinizing Probabilistic Data) so as to continue to use existing end-applications that take only deterministic input. The conflicts in the two problem settings lead to many different challenges. Authors in the paper address a problem that chooses the set of uncertain objects to be cleaned, in order to achieve the best development in the quality of query answers. However, their aim is to improve quality of single query, while our aim is to optimize quality of overall query workload.

For a given workload of triggers/queries, the significant challenge is to find the deterministic representation of the query which will efficiently optimize certain quality metrics of the answer to these triggers/queries. Addressing the problem of determinizing, a collecting items to optimize set-based quality metrics, such as F-measure. They have extended the solutions to handle a query model where mutual exclusion is present among the tags. It also shows that inter-relation among the tags can be leveraged in our solutions to get better output. They also demonstrate that the solutions are made to handle various types of queries. The empirical demonstration of the proposed models are very efficient and reach high-quality outputs that are very near to those of the optimal solution. They have also demonstrated that there is robust to small changes in the original query workload

II. Related work

A. Determinizing Probabilistic Data.

While we do not know of any previous work that directly addresses the problem of determinizing probabilistic data as studied in this paper, the works that are very related to ours are [1],[6]. They search how to determinize answers to a query over a *probabilistic* database. We are only concerned in top deterministic representation of data so as to keep on using accessible end-applications that take only deterministic input. The differences in the two problem settings lead to different challenges. Authors in [7] deal with a problem that chooses the list of uncertain objects to be cleaned, in order to realize the best development in the class of query answers. However, their aim is to get better value of single query, while ours is to optimize quality of overall query workload. Also, the focus is on how to choose the most excellent sets of objects and each chosen object is cleaned by human clarification, whereas we determinize all objects automatically. These differences effectively lead to different optimization challenges. Another allied area is MAP inference in graphical model [7], [8], whose goal is to discover the assignment to each variable that together maximizes the probability defined by the model. The determinization problem for the cost-based metric can be seen as an case of MAP inference problem. If we look the problem that way, the test in front of us is to develop a fast and high-valued inexact code to solve the equivalent NP-hard problem.

B.Probabilistic Data Models.

A range of highly developed data models have been proposed in the past. Our focus however was determinizing

ISSN: 2278 – 7798

probabilistic objects, example image tags and speech output, for which the probabilistic attribute model suffices. We observe that determining probabilistic data stored in more highly advanced probabilistic models such as tree might also be interesting and can be possible [1]. Furthermore, our work to deal with data of such high complexity is an interesting future direction of work. There are many research efforts related that deals with the problem of selecting terms to number a document for document retrieval.

C.Key term Selection

There are many research efforts related that deals with the problem of selecting terms to number a document for document retrieval. A term-centric pruning method explained in keeps topmost postings for each and every term according to the individual score impact that each and every posting will have if the term is seen in an for the function search query [1]. We propose a scalable term selection for categorization of text, which is based upon coverage of the terms coverage of the terms The focus of these research efforts is based on relevance – that is, finding the correct set of terms that are most relevant to document. In our problem, a set of possibly relevant terms and their relevance to the document are already given by other data dealing out techniques. Thus, our goal is not to find the relevance of terms to documents, but to find and select keywords from the given set of terms to represent the document, such that the quality of answers to triggers/queries is optimized.

D .Query intent disambiguation.

Query information in such type of works is used to calculate many appropriate terms for queries, of queries. However, our aim is not to guess correct terms, but to find the correct keywords from the terms that are automatically generated by automated data generation tool[1].

E.Query and tag suggestions.

Another related explore area is that of query suggestion and tag suggestion [11]–[13]. On the basis of query-flow graphical representation of query information, authors in [11] develop a measure of semantic similarity between queries, which is used for the task of producing diverse and useful recommendations. Rae *et al.* [12] introduces an extendable structure of tag suggestion, using co-incidence examination of tags used in user detailed contents such as personal, social contact, social group and non user specific contents. The main objective of this is on how to make similarities and correlations between queries/tags and recommend queries/tags based on those information. However, our aim is not to measure similarity between object tags and queries, but to select tags from a given set of uncertain tags to optimize certain quality metric of answers to multiple.

III. CONCLUSION

Hence, from this paper we have considered problem of deteminizing uncertain objects in order to organize and store such data in already existing systems example Flickr which only accepts deterministic value. Our aim is to produce a deterministic depiction that optimizes the quality of answers to queries/triggers that execute over the deterministic data representation .As in future work, we plan to perform project on efficient determinization algorithms that are orders of scale faster than the enumeration based best solution but achieves almost the same excellence as the optimal solution and search determinization techniques as per the application context, wherein users are also involved in retrieving objects in a ranked order.

IV. ACKNOWLEDGEMENT

The authors would really like to give thanks the publishers, researchers for creating their resources obtainable and academics for his or her guidance. We have a tendency to conjointly impart the faculty authority for providing the desired infrastructure support. Finally, we'd wish to extend dear feeling to friends &family members.

V. REFERENCES

[1] K. Jie Xu, Dmitri V. Kalashnikov, and Sharad Mehrotra, "Query Aware Determinization of Uncertain Objects," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no.1, Jan. 2015.

[2] 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. 1373–1387, Sept. 2011.

[3] J. Li and J. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, Sept. 2003.

ISSN: 2278 – 7798

[4] 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.

[5] B. Minescu, G. Damnati, F. Bechet, and R. de Mori, "Conditional use of word lattices, confusion networks and 1-best string hypotheses in a sequential interpretation strategy," in *Proc. ICASSP*, 2007.

[6] R. Nuray-Turan, D. V. Kalashnikov, S. Mehrotra, and Y. Yu, "Attribute and object selection queries on objects with probabilistic attributes," *ACM Trans. Database Syst.*, vol. 37, no. 1, Article 3, Feb. 2012.

[7] V. Jojic, S. Gould, and D. Koller, "Accelerated dual decompositionfor MAP inference," in *Proc. 27th ICML*, Haifa, Israel, 2010.

[8] D. Sontag, D. K. Choe, and Y. Li, "Efficiently searching for frustrated cycles in map inference," in *Proc. 28th Conf. UAI*, 2012.

[9] S. Bhatia, D. Majumdar, and P. Mitra, "Query suggestions in the absence of query logs," in *Proc. 34th Int. ACM SIGIR*, Beijing, China, 2011.

[10] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA, USA: MIT Press, 1999.

[11] I. Bordino, C. Castillo, D. Donato, and A. Gionis, "Query similarity by projecting the query-flow graph," in *Proc. 33rd Int. ACM SIGIR*, Geneva, Switzerland, 2010.

[12] A. Rae, B. Sigurbjörnsson, and R. V. Zwol, "Improving tag recommendation using social networks," in *Proc. RIAO*, Paris, France, 2010.

[13] B. Sigurbjörnsson and R. V. Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th Int. Conf. WWW*, New York, NY, USA, 2008.

ISSN: 2278 – 7798