ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Query Aware Determinization of Uncertain Objects

C.Shobana, Prem Kumar

PG Scholar, Department of CSE, Sri Venkateswara of College of Engineering and Technology, Thiruvallur, India

Assistant Professor, Department of CSE, Sri Venkateswara of College of Engineering and Technology, Thiruvallur, India

ABSTRACT: This paper considers the problem of determinizing probabilistic data to enable such data to be stored in legacy systems that accept only deterministic input. Probabilistic data may be generated by automated data analysis/enrichment techniques such as entity resolution, information extraction, and speech processing. The legacy system may correspond to pre-existing web applications such as Flickr, Picasa, etc. The goal is to generate a deterministic representation of probabilistic data that optimizes the quality of the end-application built on deterministic data. We explore such a determinization problem in the context of two different data processing tasks—triggers and selection queries. We show that approaches such as thresholding or top-1 selection traditionally used for determinization lead to suboptimal performance for such applications. Instead, we develop a query-aware strategy and show its advantages over existing solutions through a comprehensive empirical evaluation over real and synthetic datasets.

I. INTRODUCTION

With the advent of cloud computing and the proliferation of web-based applications, users often store their data in various existing web applications. Often, user data is generated automatically through a variety of signal processing, data analysis/enrichment techniques before being stored in the web applications. For example, modern cameras support vision analysis to generate tags such as indoors/outdoors, scenery, landscape/portrait, etc. Modern photo cameras often have microphones for users to speak out a descriptive sentence which is then processed by a speech recognizer to generate a set of tags to be associated with the photo.

The photo (along with the set of tags) can be streamed in real-time using wireless connectivity to Web applications such as Flickr. Pushing such data into web applications introduces a challenge since such automatically generated content is often ambiguous and may result in objects with probabilistic attributes. For instance, vision analysis may result in tags with probabilities and, likewise, automatic speech recognizer (ASR) may produce an N-best list or a confusion network of utterances. Such probabilistic data must be “determinized" before being stored in legacy web applications. We refer to the problem of mapping probabilistic data into the corresponding deterministic representation as the determinizationproblem.

Many approaches to the determinizationproblem can be designed. Two basic strategies are the Top-1 and All techniques, wherein we choose the most probable value / all the possible values of the attribute with non-zero probability, respectively. For instance, a speech recognition system that generates a single answer/tag for each utterance can be viewed as using a top-1 strategy.

Another strategy might be to choose a threshold τ and include all the attribute values with a probability higher than τ . However, such approaches being agnostic to the end-application often lead to suboptimal results as we will see later. A better approach is to design customized determinization strategies that select a determinized representation which optimizes the quality of the end-application.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

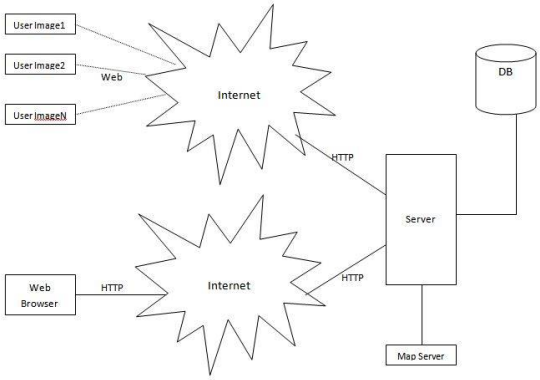
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Flickr supports effective retrieval based on photo tags. In such an application, users may be interested in choosing determinized representation that optimizes set-based quality metrics such as F-measure instead of minimizing false positives/negatives. In this paper, we study the problem of deteminizing datasets with probabilistic attributes (possibly generated by automated data analyses/enrichment). Our approach exploits a workload of triggers/queries to choose the "best" deterministic representation for two types of applications – one, that supports triggers on generated content and another that supports effective retrieval. Interestingly, the problem of determinization has not been explored extensively in the past. The most related research efforts are, which explore how to give deterministic answers to a query (e.g. conjunctive selection query.

overprobabilisitc database. Unlike the problem of determinizing an answer to a query, our goal is to determinize the data to enable it to be stored in legacy deterministic databases such that the determinized representation optimizes the expected performance of queries in the future. Solutions in cannot be straightforwardly applied to such a determinization problem.

II. SYSTEM ARCHITECTURE



EXISTING SYSTEM

- Many approaches to the determinization problem can be designed. Two basic strategies are the Top-1 and All techniques, wherein we choose the most probable value / all the possible values of the attribute with non-zero probability, respectively.
- For instance, a speech recognition system that generates a single answer/tag for each utterance can be viewed as using a top-1 strategy. Another strategy might be to choose a threshold τ and include all the attribute values with a probability higher than τ .
- Existing system works address a problem that chooses the set of uncertain objects to be cleaned, in order to achieve the best improvement in the quality of query answers.
- There are several related research efforts that deal with the problem of selecting terms to index document for document retrieval. A term-centric pruning method described in existing system retains top postings for each term according to the individual score impact that each posting would have if the term appeared in an adhoc search query.

DISADVANTAGES

- Often lead to suboptimal results.
- They explore how to determinize answers to a query over a probabilistic In contrast, we are interested in best deterministic representation of data (and not that of a answer to a query) so as to continue to use existing end-applications that take only deterministic input.
- Their goal is to improve quality of single query, while ours is to optimize quality of overall query workload.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

III. PROPOSED SYSTEM

- In this paper, we study the problem of deteminizing datasets with probabilistic attributes (possibly generated by automated data analyses/enrichment).
- Our approach exploits a workload of triggers/queries to choose the "best" deterministic representation for two types of applications – one, that supports triggers on generated content and another that supports effective retrieval.
- Interestingly, the problem of determinization has not been explored extensively 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 probabilisitic database.
- Unlike the problem of determinizing an answer to a query, our goal is to determinize the data to enable it to be stored in legacy deterministic databases such that the determinized representation optimizes the expected performance of queries in the future. Solutions cannot be straight forwardly applied to such a determinization problem.

ADVANTAGES:

- We introduce the problem of determinizing probabilistic data. Given a workload of triggers/queries, the main challenge is to find the deterministic representation of the data which would optimize certain quality metrics of the answer to these triggers/queries.
- Solves the problem of determinization by minimizing the expected cost of the answer to queries.
- We develop an efficient algorithm that reaches near-optimal quality.
- The proposed algorithms are very efficient and reach high-quality results that are very close to those of the optimal solution. We also demonstrate that they are robust to small changes in the original query workload.

IV. CONCLUSION

In this paper we have considered the problem of determinizing uncertain objects to enable such data to be stored in pre-existing systems, such as Flickr, that take only deterministic input. The goal is to generate a deterministic representation that optimize the quality of answers to queries/triggers that execute over the deterministic data representation. We have proposed efficient determinization algorithms that are orders of magnitude faster than the enumeration based optimal solution but achieves almost the same quality as the optimal solution.

FUTURE ENHANCEMENT:

As future work, we plan to explore determinization techniques in the context ofapplications, wherein users are also interested in retrieving objects in a ranked order.

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. 1373–1387, Sept. 2011.
[2] 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.
[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] B. Minescu, G. Damnati, F. Bechet, and R. de Mori, "Conditional use of word lattices, confusion networks and 1-best stringhypotheses in a sequential interpretation strategy," in *Proc. ICASSP*, 2007.
[5] R. Nuray-Turan, D. V. Kalashnikov, S. Mehrotra, and Y. Yu, "Attribute and object selection queries on objects with probabilisticattributes," *ACM Trans. Database Syst.*, vol. 37, no. 1, Article 3, Feb. 2012.
[6] J. Li and A. Deshpande, "Consensus answers for queries over probabilistic databases," in *Proc. 28th ACM SIGMOD-SIGACTSIGARTSymp. PODS*, New York, NY, USA, 2009.
[7] M. B. Ebarhimi and A. A. Ghorbani, "A novel approach for frequent phrase mining in web search engine query streams,"in *Proc. 5th Annu. Conf. CNSR*, Frederlcton, NB, Canada, 2007.
[8] S. Bhatia, D. Majumdar, and P. Mitra, "Query suggestions in the absence of query logs," in *Proc. 34th Int. ACM SIGIR*, Beijing,China, 2011.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

[9] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA, USA: MIT Press, 1999.
[10] D. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph," *ACM Trans.Database Syst.*, vol. 31, no. 2, pp. 716–767, Jun. 2006.
[11] K. Schnaitter, S. Abiteboul, T. Milo, and N. Polyzotis, "On-line index selection for shifting workloads," in *Proc. IEEE 23rd Int.Conf. Data Eng. Workshop*, Istanbul, Turkey, 2007.
[12] P. Unterbrunner, G. Giannikis, G. Alonso, D. Fauser, and D. Kossmann, "Predictable performance for unpredictable workloads," in*Proc. VLDB*, Lyon, France, 2009.
[13] R. Cheng, J. Chen, and X. Xie, "Cleaning uncertain data with quality guarantees," in *Proc. VLDB*, Auckland, New Zealand, 2008.
[14] V. Jojic, S. Gould, and D. Koller, "Accelerated dual decomposition for MAP inference," in *Proc. 27th ICML*, Haifa, Israel, 2010.
[15] D. Sontag, D. K. Choe, and Y. Li, "Efficiently searching for frustrated cycles in map inference," in *Proc. 28th Conf. UAI*, 2012.
[16] P. Andritsos, A. Fuxman, and R. J. Miller, "Clean answers overdirty databases: A probabilistic approach," in *Proc. 22nd ICDE*,Washington, DC, USA, 2006.