A SURVEY ON QUERY AWARE DETERMINATION OF UNCERTAIN OBJECTS

Mr. D. BALU,
M.Tech (CSE) from
Jagruti Institute of Engineering and Technology, Telangana State, India.

Mr. P.GIRIDHAR,
Assistant Professor, Department of Computer Science and Engineering, Jagruti Institute of Engineering and Technology, Telangana State, India.

ABSTRACT: *Considers the problem of determinizing probabilistic data and such data to be stored in legacy systems that accept only deterministic input. In this paper we survey on different authors and their proposed work. Probabilistic information is also generated by machine-controlled information from entity resolution, data extraction, and speech process. The inheritance system could correspond to pre-existing internet applications like Flickr, Picasa, etc. The goal is to get a settled illustration of probabilistic information that optimizes the standard of the end-application engineered on complete information. we have a tendency to explore such a Determinization drawback within the context of 2 completely different processing tasks triggers and choice queries. we have a tendency to show that approaches like thresholding or top-1 choice historically used for Determinization cause suboptimal performance for such applications. Instead, we have a tendency to develop a query-aware strategy and show its benefits over existing solutions through a comprehensive empirical analysis over real and artificial datasets.*

Keywords: *Determinzation, uncertain data, data quality, query workload, branch and bound algorithm.*

1. INTRODUCTION

With the appearance of cloud computing and also the proliferation of web-based applications, users typically store their knowledge in numerous existing internet applications. Often, user knowledge is generated mechanically through a range of signal process, knowledge analysis/enrichment techniques before being hold on within the internet applications. For instance, fashionable cameras support vision analysis to get tags like indoors/outdoors, scenery, landscape/portrait, etc. fashionable icon cameras typically have microphones for users to talk out a descriptive sentence that is then processed by a speech recognizer to get a collection of tags to be related to the icon. The icon (along with the set of tags) may be streamed in time period exploitation wireless property to internet applications like Flickr. Pushing such knowledge into internet applications introduces a challenge since such mechanically generated content is commonly ambiguous and will end in objects with probabilistic attributes. as an example, vision analysis could end in tags with possibilities and, likewise, automatic speech recognizer (ASR) could turn out Associate in Nursing N-best list or a confusion network of utterances. Such probabilistic knowledge should be "determinized" before being hold on in heritage internet applications. We refer to the problem of mapping probabilistic data into the corresponding deterministic representation as the determinizationproblem. 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 . 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 determination strategies that select a determinized representation which optimizes the quality of the end-application. 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. Over probabilisitc 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.

2. RELATED WORK

Many advanced probabilistic data models were used in proposed systems. Here the centre of attention however was determinizing probabilistic objects, such as speech output and image tags, for which the probabilistic attribute model meet the requirements. It is to be noted that determining probabilistic data stored in more advanced probabilistic representation such as tree structures is also used. Several related research efforts that contract with the problem of selecting terms to index document for document retrieval. A term-centric pruning method explains in keeps top postings for each term according to the individual score impact that each posting would have if the term appeared in a temporary search query. Here we propose a scalable term selection for text classification, is nothing but which is based on coverage of the terms. The centre of these research efforts is on significance – that is, getting the right set of terms that are most relevant to this paper. In our problem, a set of probably appropriate terms and their significance to the document are already specified by other data processing techniques. Thus, our objective is not to explore the significance of terms to documents, but to select keywords from the given set of terms to represent the paper, such that the quality of answers to triggers

or queries is optimized. The main advantage of our proposed system is it will resolve the problem of determinization by reducing the expected cost of the answer to queries. Here we develop an efficient algorithm that achieves near-optimal quality. The algorithms which we are advice are very capable and reach high-quality results that are very close to those of the optimal solution [11]. Cutting edge information preparing strategies, for example, substance determination, information cleaning, data extraction, and mechanized labeling frequently deliver results comprising of items whose traits may contain instability. This vulnerability is every now and again caught as an arrangement of various fundamentally unrelated quality decisions for each questionable characteristic alongside a measure of likelihood for option values. On the other hand, the lay end client, and some end-applications, won't not have the capacity to decipher the outcomes if yielded in such a structure. Along these lines, the inquiry is the manner by which to present such results to the client practically speaking, for instance, to bolster characteristic quality choice and article determination inquiries [12] the client may be keen on. Specifically, in this article we examine the issue of boosting the nature of these choice questions on top of such a probabilistic representation. The quality is measured utilizing the standard and generally utilized set-based quality measurements. We formalize the issue and after that create efficient approaches that give superb responses to these questions. Uncertain data are inherent in some important applications, such as environmental surveillance, market analysis, and quantitative economics research. Uncertain data in those applications are generally caused by factors like data randomness and incompleteness, limitations of measuring equipment, delayed data updates, etc [5]. Due to the importance of those applications and the rapidly increasing amount of uncertain data collected and accumulated, analyzing large collections of uncertain data has become an important task and has attracted more and more interest from the database community.

3. LITERATURE SURVEY

A semantics-based approach for speech annotation of images [D. V. Kalashnikov, S. Mehrotra, J. Xu, and N. Venkatasubramanian]

Associating textual annotations/tags with multimedia content is among the most effective approaches to organize and to support search over digital images and multimedia databases. Despite advances in multimedia analysis, effective tagging remains largely a manual process wherein users add descriptive tags by hand, usually when uploading or browsing the collection, much after the pictures have been taken. This approach, however, is not convenient in all situations or for many applications, e.g., when users would like to publish and share pictures with others in real time. An alternate approach is to instead utilize a speech interface using which users may specify image tags that can be transcribed into textual annotations by employing automated speech recognizers. Such a speech-based approach has all the benefits of human tagging without the

cumbersomeness and impracticality typically associated with human tagging in real time. The key challenge in such an approach is the potential low recognition quality of the state-of-the-art recognizers, especially, in noisy environments. In this paper, we explore how semantic knowledge in the form of co-occurrence between image tags can be exploited to boost the quality of speech recognition. We postulate the problem of speech annotation as that of disambiguating among multiple alternatives offered by the recognizer. An empirical evaluation has been conducted over both real speech recognizer's output as well as synthetic data sets. The results demonstrate significant advantages of the proposed approach compared to the recognizer's output under varying conditions.

Automatic linguistic indexing of pictures by a statistical modeling approach [J. Li and J. Wang]

Automatic linguistic indexing of pictures is an important but highly challenging problem for researchers in computer vision and content-based image retrieval. In this paper, we introduce a statistical modeling approach to this problem. Categorized images are used to train a dictionary of hundreds of statistical models each representing a concept. Images of any given concept are regarded as instances of a stochastic process that characterizes the concept. To measure the extent of association between an image and the textual description of a concept, the likelihood of the occurrence of the image based on the characterizing stochastic process is computed. A high likelihood indicates a strong association. In our experimental implementation, we focus on a particular group of stochastic processes, that is, the two-dimensional multiresolution hidden Markov models (2D MHMMs). We implemented and tested our ALIP (Automatic Linguistic Indexing of Pictures) system on a photographic image database of 600 different concepts, each with about 40 training images. The system is evaluated quantitatively using more than 4,600 images outside the training database and compared with a random annotation scheme. Experiments have demonstrated the good accuracy of the system and its high potential in linguistic indexing of photographic images.

Image annotation refinement using random walk with restarts [C. Wangand, F. Jing, L. Zhang, and H. Zhang]

Image annotation plays an important role in image retrieval and management. However, the results of the state-of-the-art image annotation methods are often unsatisfactory. Therefore, it is necessary to refine the imprecise annotations obtained by existing annotation methods. In this paper, a novel approach to automatically refine the original annotations of images is proposed. On the one hand, for Web images, textual information, e.g. file name and surrounding text, is used to retrieve a set of candidate annotations. On the other hand, for non-Web images that are lack of textual information, a relevance model-based algorithm using visual information is used to decide the candidate annotations. Then, candidate annotations are re-ranked and only the top ones are reserved as the final annotations. To re-rank the annotations, an algorithm

using Random Walk with Restarts (RWR) is proposed to leverage both the corpus information and the original confidence information of the annotations. Experimental results on both non-Web images of Corel dataset and Web images of photo forum sites demonstrate the effectiveness of the proposed method.

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

Within the context of a deployed spoken dialog service, this study presents a new interpretation strategy based on the sequential use of different ASR output representations: 1-best strings, word lattices and confusion networks. The goal is to reject as early as possible in the decoding process the nonrelevant messages containing non-speech or out-of-domain content. This is done through the 1-pass of the ASR decoding process thanks to specific acoustic and language models. A confusion network (CN) is then calculated for the remaining messages and another rejection process is applied with the confidence measures obtained in the CN. The messages kept at this stage are considered relevant; therefore the search for the best interpretation is applied to a richer search space than just the 1-best word string: either the whole CN or the whole word lattice. An improved, SLU oriented, CN generation algorithm is also proposed that significantly reduces the size of the CN obtained while improving the recognition performance. This strategy is evaluated on a large corpus of real users ' messages obtained from a deployed service.

Attribute and object selection queries on objects with probabilistic attributes [R. Nuray-Turan, D. V. Kalashnikov, S. Mehrotra, and Y. Yu]

Modern data processing techniques such as entity resolution, data cleaning, information extraction, and automated tagging often produce results consisting of objects whose attributes may contain uncertainty. This uncertainty is frequently captured in the form of a set of multiple mutually exclusive value choices for each uncertain attribute along with a measure of probability for alternative values. However, the lay end-user, as well as some end-applications, might not be able to interpret the results if outputted in such a form. Thus, the question is how to present such results to the user in practice, for example, to support *attribute-value selection* and *object selection* queries the user might be interested in. Specifically, in this article we study the problem of maximizing the quality of these selection queries on top of such a probabilistic representation. The quality is measured using the standard and commonly used set-based quality metrics. We formalize the problem and then develop efficient approaches that provide high-quality answers for these queries. The comprehensive empirical evaluation over three different domains demonstrates the advantage of our approach over existing techniques.

4. CONCLUSIONS

In this paper we have considered problem of determinizing uncertain objects in order to organize and store such data in already existing systems example Flickr which only accepts deterministic value. Our goal to produce a deterministic illustration 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.

REFERENCES:

[1] Jie Xu, Sharad Mehrotra," Query Aware Determinization of Uncertain Objects" ,IEEE Transactions on knowledge and data engineering, VOL. 27, NO. 1, January 2015.

[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 string hypotheses in a sequential interpretation strategy," in Proc. ICASSP, 2007.

[5] Jian Pei, Ming Hua," Query Answering Techniques on Uncertain and Probabilistic Data" In VLDB, pages 1151- 1154, 2006.

[6] Umesh Gorela1, Bidita Hazarika2, Abhinesh Tiwari3, Priti Mithari," Survey on Query Aware Strategy for Determining Uncertain Probabilistic Data", in (IJSETR), Volume 4, Issue 10, October 2015 3510

[7] 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.

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

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

[10] 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.

[11] P.Jhancy, K.Lakshmi ,Dr.S.Prem Kumar," Query Aware Determinization of Uncertain Objects" in ijcert Volume

2, Issue 12, December-2015, pp. 904-907.

[12] 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.

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

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

[15] D. Carmel et al., "Static index pruning for information retrieval systems," in Proc. 24th Annu. Int. ACM SIGIR, New Orleans, LA, USA, 2001.