Query Aware Uncertain Objects Determinization Along with Tag Ranking

Kazi S.C, Amrit Priyadarshi

Abstract— **Some automatic techniques are present which can generate probabilistic data with aid of entity resolution, speech processing, information extraction and many more. This paper deals with the problem of determinizing the uncertainty of data is analyzed to store in the legacy systems which accepts only deterministic input. Some web applications as Flicker, Picasa, etc corresponds to the legacy system. The main target is to generate deterministic presentation of probabilistic data to optimize quality of application built on deterministic data and to give proper tag ranking to the objects for easy retrieval to user. Two data processing methods as trigger and selection queries explores the determinization issue. To tackle the problem and enhance the performance of applications a query-aware strategy can be implemented. We use branch and bound algorithm to find optimal solution and threshold-cut approach.**

Index Terms— **Determinization, Uncertain data, query workload, branch and bound algorithm, Tag ranking**

I. INTRODUCTION

The advent of cloud computing and web based application changed the way of data storage. Various signal processing, data analysis techniques automatically generate. User data and get stored in the web application. Such as modern cameras supporting vision analysis and speech recognizer to generate automated tags along the pictures[2].Automated generated contents are usually ambiguous and may pose some challenges of probabilistic attributes if streamed in web based applications like Flicker, Picasa. Vision analysis generated tags results in probabilities[3],[4] and the speech recognizer ,viz., Automated Speech Recognizer(ASR),produces N-best list or confusion of utterances [2],[5]. Determinization of such Probabilistic data must be done before storing on legacy web based applications. The mapping problem of probabilistic data into deterministic representation termed as Determinization Problem.

Among many approaches two basic strategies used for determinization problem are Top-1 and All techniques. In these techniques most probable value or all the possible values of attributes with non-zero probability are chosen. To optimize the End –application quality customized determinization strategies is designed. The paper deals with determinizing datasets with probabilistic attributes. Study exploits workload of queries/triggers to select "best" deterministic representation for two applications (1) that supports effective retrieval and (2) that supports triggers on generated content. The problem of determinization can be solved by minimizing the expected cost of answer to queries. Branch and Bound algorithm developed to find an approximate optimal solution, the problem of determinizing a collection of objects addressed to optimize the set based quality metrics as F-measures.

The tags associated with an image are generally in a random order with no any importance or relevance information, hence limiting the effectiveness of the tags in search and other applications. The tags associated with Flickr images are almost order less, which limits the effectiveness of the tags in many related applications. The paper proposes an approach to rank the tags for each image according to their relevance levels. The tag-based search, group recommendation and tag recommendation are the three approaches proposed by tag ranking. Tag ranking can help to organize and retrieve the large set of images as per the convenience and importance of relevance. Through a probabilistic approach the relevance score of individual tag can be estimated provided an image and its tag information. The Kernel Density Estimation (KDE) is used for determining the relevance score of tags. Then a random walk-based refinement is done to check the tag ranking performance by comparing tag relationships.

II. RELATED WORK

J. Li,and J. Wang.[3] proposed a statistical modeling approach to automatic linguistic indexing problem of pictures. Categorized images are used to train statistical models representing concepts. The extent of association can be measured among the textual description and a image. Here main focus is on specific group of stochastic processes. It gives accuracy of system with high potential in indexing of photographic images.

A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo.[4] introduces methodology for click through logs, content of search engine and query specific information data results pages to study query intent characteristics especially commercial ones Improvement to search results and its personalization are initiated by understanding user query intents. The number of displayed ads on click through and the effect of query are studied. For effective detection of queries the findings of study as click through features, content of search results and query features are together effective. Modeling query intent can improve the accuracy of predicting ads click through for any unseen query

C. Wangand, F. Jing, L. Zhang, and H. Zhang.[5]
propose the image retrieval and management. It presents the approach for refinement of the annotations of images. Textual information data is used to retrieve a candidate set. A relevance model algorithm is used for non-web based images

Manuscript received July 05, 2016
Kazi S.C., Computer Engineering Department, DGOI FOE,Swami-Chincholi, Bhigwan, Pune, India
Amrit Priyadarshi, Computer Engineering Department, DGOI FOE,Swami-Chincholi, Bhigwan,Pune,India

to decide candidate annotation. Re-ranking is done for candidate annotation and the final annotations are the top ones from these. Random walk with restart (RWR) algorithm is used for re-ranking of annotations to leverage the original confidence information and the corpus data of the annotation. Photo forum sites demonstrates the effectiveness of Corel dataset and web images of photo.

R. Nuray-Turan, D. V. Kalashnikov, S. Mehrotra, and Y. Yu[7] focus on the problem of maximization the quality of selection on top queries with probabilistic representation. The set-based quality metric is used to measure quality. Attributes with object uncertainty are produced due to some recent techniques such as entity resolution, information extortion, data cleansing, and automated techniques. This uncertainty can be captured in form of set of multiple mutually exclusive values for each uncertain attribute with probability for each attributes. The high quality answers are developed for such queries. The effective evaluation over 3 domains demonstrates advantage of this approach along the existing one.

.

B. Sigurbjörnsson and R. V. Zwol [8] investigate tagging phases to assist users. Users can share online photos with friends, family, social network by using available services such as Flicker and Zoomer at a large rate. Users manually can annotate these photos using their tags, which describes contents of additional semantically and contextual information. The papers contributions are twofold, that is, present the results by means of a tagging characterization and analyze a representative snapshot of Flicker focus is on how users tag a photo and what information is considered along it. Hence based on all these analysis, paper presents and evaluates tag recommendations strategies which support the users to annotate photos by providing a set of tags which can be used by users to add to their photo. The empirical evaluation results show the effectiveness of tag recommendation for a variety of photos having different levels of originality or genuineness.

III. PROBLEM STATEMENT

The paper deals with the problem of determinizing the uncertainty of data is analyzed to store in the legacy systems which accept only deterministic input. The main objective of paper is to generate deterministic presentation of probabilistic data to optimize quality of application built on deterministic data and to give proper tag ranking to the objects for easy retrieval to user. This Probabilistic data can be generated by automated data analysis/enrichment techniques such as entity resolution, information extraction, and speech processing.

IV. PROPOSED SYSTEM

The paper deals with the problem of determinizing datasets with probabilistic attributes generated by automated data analyses or enrichment. The approach uses queries to choose deterministic representation for two types of applications (1) that supports effective retrieval, (2) that supports triggers on generated content. The paper provides answer to deterministic query over probabilistic database. In future also the answer to the deterministic database which is stored in the legacy system can optimize these query performance with determinizing an answer to a query. Solutions cannot be straightly applied to such a determinization problem; hence the use of Branch and Bound algorithm is applied.

The proposed system architecture consists of following modules:

A. Data owner Module
B. Data User Module
C. Query-Level Optimization
D. Query Work Load
E. Tag Ranking

A) Data owner Module

The first module developed is the data owner module. It is not feasible when the tags numbers are large, which is the case in multiple data sets. Hence, a branch-and-bound algorithm is used to solve EDCM approximately. The approximate solution meets the quality that matches exact one, while being orders of magnitude more efficient is demonstrated. First, in this module, Data Owner New user should register then only the data owner can Login to Application. After the Registration is completed, then admin approved user only to login. Then only user can login to home page. User first search the image file. The use of any queries that processed to get output in Grid view can be done. Grid view shows related search items. User selects a particular image and Download it. Also user can modify and Update their Details. Home Page has about the project Details.

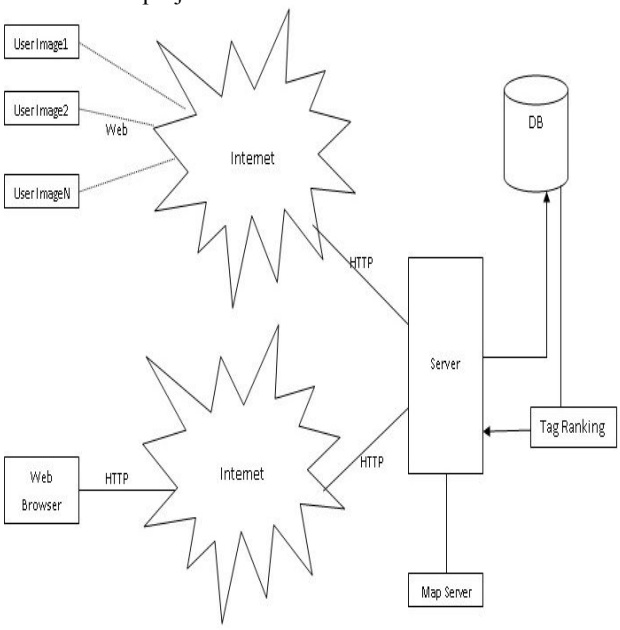

Fig. 1 System Architecture

B) Data User Module

In this module, the admin provides the functionality of verifying registration and approving them, then uploading the images and providing different tags to it. Admin can view user details and lock particular user. In this the Chart view shows user performance, execution time and number of images downloaded by user.

C) Query-Level Optimization

The Query–Level optimization helps improve the performance of the Branch and Bound algorithm. The expected cost of particular queries can be exactly determined if sequence of nodes is provided. Branch and Bound method

can provide faster output over brute-force enumeration. The used approach finds solution sets in a greedy fashion with lower cost solutions discovered firstly. The partial tag selection decides whether to include any tag in solution set or not, and here ever tag represents a node. The defined sequence helps in tag selection using the concept of defined sequence.

D) Query Work Load

Query Workload is Set of Queries executing over the deterministic representation based on the past querying pattern, or corpus of data. The main approach is to exploits a workload of uses queries to choose deterministic representation for two types of applications (1) that supports effective retrieval,(2)that supports triggers on generated content. The solution is tested over .any synthetic query workloads where parameters of workloads follow specific distributions.

E) Tag Ranking

The paper proposes a tag ranking scheme, designed to automatically rank the tags as per the relevance to image content. For this initial relevance scores is to be estimated for the tags which are based on probability density estimation, and then the relevance scores are refined by performing a random walk over a graph of tag similarity. Tag ranking can be applied into three applications: (1)tag recommendation, (2)group recommendation, and (3) tag-based image search, which demonstrates that the proposed tag ranking approach really boosts the performances of social-tagging related applications.

V. RESULT ANALYSIS

Query Aware searched the objects based on tags in randomized order using Branch and Bound technique. It is improved in this paper by extending Tag Ranking concept

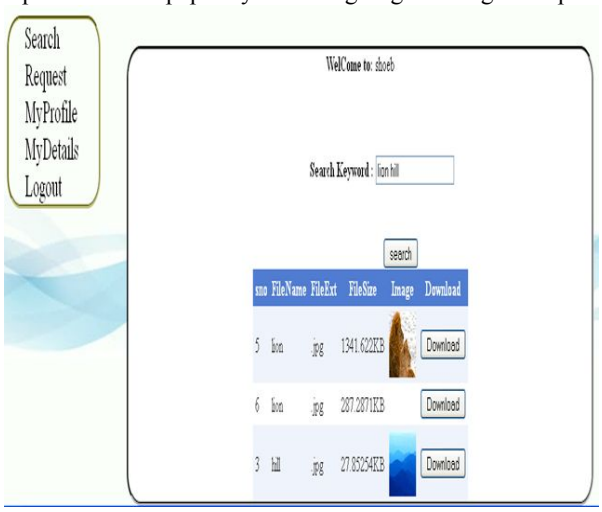

Fig. 2 Images retrieved based on query sent by user

CONCLUSION

In this paper the problem of uncertain objects determinization is studied so as to store them in some pre-existing system applications like Flicker. The main objective of this is to optimize quality of solutions to queries/trigger which execute over deterministic data representation. Over Existing methods the proposed paper provides us with efficient algorithms. Tag ranking helps for easy retrieval of objects to user. For future work new opportunities and facilities in social media can be developed by tagging services.

ACKNOWLEDGEMENT

I would like to thank all people who help me in different way. Especially, I am thankful to my guide and P.G. Co-ordinator Prof. AmritPriyadarshi for his continuous support and guidance in my work. Also, I would like thank H.O.D. of Computer Engineering Department, Prof. S. S. Bere for motivating me. Lastly, I thank to IJERT who have given opportunity to present my paper.

REFERENCES

[1] JieXu, Dmitri V. Kalashnikov, and SharadMehrotra, "Query Aware Determinization of Uncertain Objects"Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 1, JANUARY 2015

[2] D. V. Kalashnikov,S. Mehrotra,J. Xu,and N. Venkatasubramanian, " A semantic-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.

[4] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo, "Classifying and characterizing query intent," in Proc. 31th ECIR, Toulouse, France, 2009.

[5] 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.

[6] 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.

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

[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] A.Anagnostopoulos, L. Becchetti, C. Castillo, and A. Gionis, "An optimization framework for query recommendation," in Proc. 3rd ACM Int. Conf. WSDM, New York, NY, USA, 2010.

[11] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, Hong-Jiang Zhang, "Tag Ranking", Beijing, 100190, P.R.China, 2009