International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume2, Issue 6 (June '2016)

SEARCH ENGINE BASED OBJECT SEARCH PATTERN APPLICATION USING CATEGORY CLASSIFICATION

M.Niveditha ,
PG Student,
Sasurie Academy of Engineering,
Coimbatore, India.

K.Sivachandran,
Assistant Professor,
Sasurie Academy of Engineering,
Coimbatore, India.

Abstract: This paper explains the problem of determinizing probabilistic data which takes only un deterministic input. While accepting those types of un deterministic input data the legacy systems stores those type of data. The automated data analysis/enrichment techniques generate probabilistic data. These techniques are entity resolution, information extraction, and speech processing. The legacy application system may correspond to the existing web applications like Flicker, Picasa, etc. The goal of this survey is to generate a deterministic representation of probabilistic data which optimizes the deterministic data quality on the end-application. Those determinization problems should be deployed in the area of two different data processing tasks they are triggers and selection queries. The top-1 selection techniques are traditionally used for determinization problem which create suboptimal performance for many applications. Instead this survey proposes a query-aware strategy and explains its advantages over existing solutions through a comprehensive empirical study over real and synthetic datasets. We propose a probabilistic topic model for analyzing and extracting content-related annotations from noisy annotated discrete data such as web pages stored in social bookmarking services. In these services, since users can attach annotations freely, some annotations do not describe the semantics of the content, thus they are noisy, i.e. not content-related. The extraction of content-related annotations can be used as a preprocessing step in machine learning tasks such as text classification and image recognition, or can improve information retrieval performance. The proposed model is a generative model for content and annotations, in which the annotations are assumed to originate either from topics that generated the content or from a general distribution unrelated to the content. We demonstrate the effectiveness of the proposed method by using synthetic data and real social annotation data for text and images.

Keywords: Determination, uncertain data, data quality, query workload, branch and bound algorithm

I.INTRODUCTION

Due to the advancement of cloud computing and the propagation of web-based applications, users need to store their data in various existing web applications. Frequently, user data is generated frequently through a variety of process namely signals processing, data analysis/enrichment techniques. This process could be done before being stored in the web applications. For example recent cameras support vision analysis to create tags such as indoors/outdoors, scenery, landscape/portrait, etc. Also these cameras often have microphones for users to speak out a descriptive sentence. This type of output should be processed by a speech recognizer to generate a set of tags to be associated with the photo [1]. The photo along with the set of tags/sub words should be streamed in real-time using wireless connectivity to Web applications. This method could be used in application like Flicker. Inserting those data into web applications introduces a new challenge. The big drawback is such type of automatically generated content is often confusing the people and also produces objects with probabilistic attributes. For instance vision analysis may produce the result with tags [2], [3].Likewise Automatic Speech Recognizer (ASR) may create an N-best list or a confusion network of utterances [1], [4]. Those types of probabilistic data must be "determinized" before being stored in legacy web applications. The determinization problem should be referred as the problem of mapping probabilistic data into the corresponding deterministic representation.

Several approaches should be followed for a *determinization* problem. Two basic strategies namely most probable value / all the possible values of the attribute with non-zero probability could be used respectively. These are called the Top-1 techniques. For example a speech recognition system that generates a single answer/tag for each utterance can be viewed as using a top-1 strategy. Another method might be to choose a threshold τ and include all the attribute values with a probability higher than τ. Those types of approaches are being doubter to the end application because they often lead to suboptimal results. A better approach among those is design customized determinization strategies which select a determinate representation which should optimize the quality of the end-application. For instance some type of end application supports triggers/alerts on a automatically generated content. Examples of those type of end-application includes publish/subscribe systems such as Google Alert, wherein users specify their subscriptions in the form of keywords (e.g.,

"California earthquake") and predicates over metadata (e.g., data type is video). Google Alert should forward all matching data items to the user based on the subscriptions details.

While considering a video about California Earthquake that is to be published on Youtube . The video contains a more number of tags that should be extracted using either automated vision processing and/or information extraction techniques applied over transcribed speech. Those type of automated tools may produce tags with probabilities like "California": 0.9, "earthquake":0.6, "election": 0.7, while the true tags of the video could be "California" and "earthquake". The determinization process useful to associate the video with appropriate tags. The subscribers who are really interested in the particular video Example "California Earthquake" is notified while the other is not frustrated by irrelevant data. Thus, in the example above, the determinization process should be used to minimize metrics such as false positives and false negatives that result from a determinized representation of data. In a Flicker application photos are uploaded automatically from cameras along with tags which may be generated based on speech annotation or image analysis. Flicker supports effective retrieval based on a photo tags. In that type of application users may be interested in choosing determinate representation which optimizes set-based quality metrics such as F-measure instead of minimizing the false positives/negatives rates.

II.LITERATURE SURVEY

In 2012 R. Nuray-Turaen et.al proposes [5] and J. Li proposes [6] which explain giving deterministic answers to a query e.g. conjunctive selection query [5] over a probabilistic database. Unlike the problem of determinizing an answer to a query the big goal of this paper is to determine the data to enable and it should be stored in legacy deterministic databases such that the determinized representation optimizes the expected performance of queries in the future.

Advantage:

- They explore how to determinate answers to a query over a *probabilistic* database.

Disadvantage:-

- Solutions in [5], [6] cannot be straightforwardly applied to such a determinization problem.
- The people should not aware of any prior work that directly addresses the issue of determinizing probabilistic data
- Most people are interested in best deterministic representation of data and not that of an answer to a query
- Existing end-applications that take only deterministic input which may not be possible in [5][6].
- These problems settings lead to different challenges.

Determinizing Probabilistic Data:

In 2008 R. Cheng et.al [13] address a problem which choose the set of uncertain objects to be cleaned, in order to achieve the best improvement in the quality of query answers. However the goal is to improve quality of single query and optimize quality of overall query workload. Also this survey focuses on how to select the best set of objects and each selected object is cleaned by using human clarification. In this all objects should be determinate automatically. These differences essentially lead to different optimization challenges.

In 2010 V. Jojic et.al proposes [14] and In 2012 D. Sontag proposes [15] .Both concentrate in the area MAP inference in graphical model which aims to find the assignment to each variable which should be jointly maximizes the probability defined by the model. The cost-based metric determinization problem can be potentially viewed as an instance of MAP inference problem. In this way the challenge leads to the developing fast and high-quality approximate algorithm for solving the corresponding NP-hard problem.

Probabilistic Data Models:

In 2008 L. Antova proposes [16] this should be a variety of advanced probabilistic data models. This survey focuses on determinizing probabilistic objects, such as image tags and speech output, for which the probabilistic attribute model suffices. The determining probabilistic data stored in more advanced probabilistic models such.

Disadvantage:-

- Extending our work to deal with data of such complexity remains difficult.

Index term selection:

In 2001 D. Carmel *et al* proposes [19]"Static index pruning for information retrieval systems," which deal with the problem of selecting terms to index document for document retrieval. A term-centric pruning method contains top postings for each item according to the individual score. Each item should have posting term appeared in an ad hoc search query.

In 2007 J. Li and M. Sun et.al proposes [20] "Scalable term selection for text categorization," propose a scalable item selection for text categorization. It should be based on coverage of the items. This paper concentrate on getting the right set of terms which is most relevant to document. In this problem, a set of possibly relevant terms and their relationship to the document
is already given by other data processing techniques. The main goal is does not explore the relevance of terms to documents, but to select keywords from the given set of terms to represent the document, such that the quality of answers to *triggers/queries* is optimized
.

Query intent disambiguation:

Reference from [21]–[24] describe History of data in which a query workload and click graph have been used. In the research area of query intent disambiguation Query logs are used to

predict more relevant terms for queries, or in other words more accurate intent of queries. The main goal is not to predict appropriate terms. In order to get the right keywords from the terms those are already created by automated data generation tool.Query workload in this setting is not a source of feature for classification but most useful to drive the optimization of the end performance of application. Thus techniques in these works are not directly applicable in the context of the problem domain we are addressing.

Query and tag recommendation:

[25] In 2010 I. Bordino et.al proposes query recommendation and tag recommendation]. Based on a query-flow graph representation of query logs a measure of semantic similarity between queries models should be developed which is used for the task of producing diverse and useful recommendations.

In 2010 Rae *et al.* [27] propose an extendable framework of tag recommendation, using co-occurrence analysis of tags used in both user specific context (personal, social contact, social group) and non user specific context (collective). The focus of these works is on how to model similarities and correlations between queries/tags and recommend queries/tags based on those information. However, our goal 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 queries.

III.EXISTING SYSTEM

This paper studies the problem of deteminizing datasets with probabilistic attributes which should be generated by automated data analyses/enrichment techniques. This approach explodes a workload of triggers/queries to choose the "best" deterministic representation for two types of application. First type supports triggers on generated content and another that supports effective retrieval. This survey introduces the solution for 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

Advantages of Existing System

- This framework solves the problem of determinization by minimizing the expected cost of the answer to queries.
- The developed a branch and-bound algorithm that finds an approximate near-optimal solution to the resulting NP-hard problem.
- Produce a collection of objects to optimize set-based quality metrics, such as F-measure.
- We develop an efficient algorithm that reaches near-optimal quality
- The solutions should be extended to handle a data model where mutual exclusion exists among tags.
- The correlations among tags can be leveraged in these solutions to get better results.
- This proposes systems are designed to handle various types of queries.
- The proposed techniques are very efficient and reach high-quality results that are very close to optimal solution.
- They are robust to small changes in the original query workload.

Disadvantages of Existing System

- Content-unrelated annotations may also constitute noise in information retrieval.
- User can attach annotations which are unrelated content, so while viewing is show unrelated content.

IV.PROPOSED SYSTEM

In this paper, we propose a probabilistic topic model for analyzing and extracting content-related annotations from noisy annotated data. The extraction of content-related annotations can improve performance of machine learning and information retrieval tasks. The proposed model can also be used for the automatic generation of content-related annotations. The proposed method is based on topic models. A topic model is a hierarchical probabilistic model, in which a document is modeled as a mixture of topics, and where a topic is modeled as a probability distribution over words. Topic models are successfully used for a wide variety of applications including information retrieval, collaborative filtering, and visualization as well as for modeling annotated data

Advantages of Proposed System

- This paper proposes a generative model for topics and annotations that takes into account relevance/irrelevance of the annotations
- We can change the unrelated content into related content by using topic model.

V.RESULTS AND DISCUSSIONS

Capturing Feedback Sessions:

Sessions for a web search is a series of successive queries to satisfy a single information need and some clicked search results. Here, feedback session consists of both clicked and un clicked URL's and ends with the last URL that was clicked in a single session. Clicked URL's state what users require and un clicked URL's reflect what users do not care about. For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze search results or clicked URL's directly because there are different feedback sessions in user click-through logs. It is unsuitable to directly use feedback sessions for inferring user search goals, because they vary a lot for different click-through logs and queries. We can represent feedback sessions by binary vector method. In this method, 0 represents unclicked URL's in click sequence and 1 represents

clicked URL's. But, binary vector representations are not informative enough.

Building pseudo-documents:
So, we used pseudo-documents to infer user search goals. Users have some unclear words for representing the interests. They use these keywords to determine whether a document can satisfy their needs. These keywords are known as "goal texts". Goal texts can reflect user information needs; they are hidden and not expressed explicitly. So, pseudo-documents are used as surrogates to approximate goal texts.
Building pseudo-documents includes two steps

- Representing the URL's in feedback session.
- Forming pseudo-documents based on URL representations:

Clustering pseudo-documents:
Pseudo-documents are clustered by using Fuzzy Keyword Search algorithm. And the optimal value will be determined through the evaluation. After clustering of all pseudo-documents, each cluster is considered as one user search goal.

Restructuring based on web search results:
Restructuring web search results is an application of inferring user search goals. Inferred user search goals are represented by the vectors and each URL's feature representation is calculated. Then, we can categorize each URL into cluster. This is performed by choosing smallest distance between URL vector and user search goal vectors and the user search goals are restructured. Possible evaluation criteria are Average precision (AP). It evaluates according to user implicit feedbacks. It is computed at the point of each relevant document in ranked sequence.

VI.CONCLUSION

We have proposed a topic model for extracting content-related annotations from noisy annotated data. We have confirmed experimentally that the proposed method can extract content-related annotations appropriately, and can be used for analyzing social annotation data. As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper, we propose a novel approach for user search goals using feedback session and pseudo document. First we construct a feedback session to analysis the user search goal from the Weblog record. It cannot provide the accurate result. So that we can introduce the pseudo document to provide the accurate results. Based on the pseudo document we have to restructure the Web search results.

VII.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 string hypotheses 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 probabilistic attributes," *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-SIGACTSIGART Symp. PODS*, New York, NY, USA, 2009.

[7] R. Cheng, J. Chen, and X. Xie, "Cleaning uncertain data with quality guarantees," in *Proc. VLDB*, Auckland, New Zealand, 2008.

[8] V. Jojic, S. Gould, and D. Koller, "Accelerated dual decomposition for 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] P. Andritsos, A. Fuxman, and R. J. Miller, "Clean answers over dirty databases: A probabilistic approach," in *Proc. 22nd ICDE*,Washington,

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

[12] J. Li and M. Sun, "Scalable term selection for text categorization," in *Proc. EMNLP-CoNLL*, Prague, Czech Republic, 2007.

[13] X. Li, Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in *Proc. 31st Annu. Int. ACM SIGIR*,Singapore, 2008.

[14] E. Pitler and K. Church, "Using word-sense disambiguation methods to classify web queries by intent," in *Proc. Conf. EMNLP*, Singapore, 2009.

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

[16] J. Teevan, S. Dumais, and D. Liebling, "To personalize or not to personalize: Modeling queries with variation in user intent," in *Proc. SIGIR*, Singapore, 2008.

[17] 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.

[18] 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.

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

