Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhanced Query Workload System on Annotated Dataset

Roopam Chaturbedi
M.Tech (CSE)
LNCT, Bhopal, India

Shweta Srivastava
Assistant Professor
LNCT, Bhopal, India

Dr. Vineet Richarya
H.O.D of Computer Department
LNCT, Bhopal, India

Abstract: ***A large number of organizations today generate and share textual descriptions of their products, services, and actions .Such collections of textual data contain significant amount of structured information, which remains buried in the unstructured text. While information extraction algorithms facilitate the extraction of structured relations, they are often expensive and inaccurate, especially when operating on top of text that does not contain any instances of the targeted structured information. We present a novel alternative approach that facilitates the generation of the structured metadata by identifying documents that are likely to contain information of interest and this information is going to be subsequently useful for querying the database. Our approach relies on the idea that humans are more likely to add the necessary metadata during creation time, if prompted by the interface; or that it is much easier for humans (and/or algorithms) to identify the metadata when such information actually exists in the document, instead of naively prompting users to fill in forms with information that is not available in the document. As a major contribution of this paper, we present algorithms that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload. Our experimental evaluation shows that our approach generates superior results compared to approaches that rely only on the textual content or only on the query workload, to identify attributes of interest.***

Keywords: Document Annotation, Query workload, searching over dataset, data extraction.

I. INTRODUCTION

In the recent years, a huge amount of data is being gathered and stored in databases everywhere across the globe, which is mainly coming from information industry and social networking sites**.** There is a need to extract and classify useful information and knowledge from such a data collected. Data mining is an interdisciplinary field of computer science and is referred to as extracting or mining knowledge from large databases. It is the process of performing automated extraction and generating the predictive information from a large database. It is actually the Process of finding the hidden information or patterns from the repositories .The fields that use Data mining techniques include medical research, marketing, telecommunication, and stock markets, health care and so on.

Data mining consists of the various technical approaches including machine learning, statistics, database system etc. The goal of the data mining process is to discover knowledge from large databases and transform into a human understandable format. The DM and knowledge discovery are essential components to the organization due to its decision making strategy. Classification, regression and clustering are three approaches of data mining in which instances are grouped into identified classes. Classification is a popular task in data mining especially in knowledge discovery and future plan. It provides the intelligent decision making. Classification not only studies and examines the existing sample data but also predicts the future behaviour of that sample data. It maps the data into the predefined class and groups. It is used to predict group membership for data instances. In Classification, the problem includes two phases first is the learning process phase in which for analysis of training data, the rule and pattern are created. The second phase tests the data and archives the accuracy of classification patterns.

There are many application domains where users create and share information; for instance, news blogs, scientific networks, social networking groups, or disaster management networks. Current information sharing tools, like content management software (e.g., Microsoft Share-

Point), allow users to share documents and annotate (tag) them in an ad hoc way. Similarly, Google Base allows users to define attributes for their

Objects or choose from predefined templates. This annotation process can facilitate subsequent information discovery. Many annotation systems allow only "untyped" keyword annotation: for instance, a user may annotate a weather report using a tag such as "Storm Category 3." Annotation strategies that use attribute-value pairs are generally more expressive, as they can contain more information than untyped approaches.

In such settings, the above information can be entered as (Storm Category, 3). A recent line of work toward using more expressive queries that leverage such annotations, is the "pay-as-you-go" querying strategy in Data spaces: In Data spaces, users provide data integration hints at query time. The assumption in such systems is that the data sources already contain structured information and the problem is to match the query attributes with the source attributes. Many systems, though, do not even have the basic "attribute-value" annotation that would make a "pay-as you- go" querying feasible. Annotations that use "attribute value" pairs require users to be more principled in their annotation efforts.

Users should know the underlying schema and field types to use; they should also know when to use each of these fields. With schemas that often have tens or even hundreds of available fields to fill, this task becomes complicated and cumbersome. This results in data entry users ignoring such annotation capabilities. Even if the system allows users to arbitrarily annotate the data with such attribute-value pairs, the users are often unwilling to perform this task: The task not only requires considerable effort but it also has unclear usefulness for subsequent searches in the future: who is going to use an arbitrary, undefined in a common schema, attribute type for future searches? But even when using a predetermined schema, when there are tens of potential fields that can be used, which of these fields are going to be useful for searching the database in the future?

Such difficulties results in very basic annotations, if any at all, that are often limited to simple keywords. Such simple annotations make the analysis and querying of the data cumbersome. Users are often limited to plain keyword searches, or have access to very basic annotation fields, such as "creation date" and "owner of document."

.

II. LITERATURE REVIEW

In this paper**[1]** the writer have proposed a paper Pay-as-You-Go User Feedback for Data space Systems This system propose a system which is a line of work towards using more expressive queries that leverage annotations is the "pay-as – you – go " querying strategy in data spaces. In data spaces users provide data integration hints at querying time. But in this paper it is assumed that data sources already contain structured information and the problem is to match the query attributes with the source attribute.

In this paper **[2]** author proposed a paper "Towards a Business Continuity Information Network for Rapid Disaster Recovery In this paper they consider the Crisis Management and Disaster Recovery have gained immense importance in the wake of recent man and nature inflicted calamities. They proposed a solution or model for pre-disaster preparation and post disaster business continuity/rapid recovery. In case of disaster need of rapid information retrieval and sharing increases. This paper proposed a disaster management model which works well at some extent but it is not considering the effective retrieval.

Here in paper **[3]** author proposed a paper "Unanimity and Compromise among Probability Forecasters" In this paper they work on probabilities of particular uncertain event. This helps us to find out annotation and attributes.

In this paper **[5]** propose a paper Random K-Label sets: An Ensemble Method for Multilevel Classification. This paper proposes an ensemble method for multilevel classification. The Random k-label sets (RAKEL) algorithm constructs each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the power set of this subset. In this way, the proposed algorithm aims to take into account label correlations using single-label classifiers that are applied on subtasks with manageable number of labels and adequate number of examples per label. Using this we can take into account the correlation between tags for annotations. But in this collaborative annotation is missing.

Here at paper **[10]** author proposed "Automatic Pattern-Taxonomy Extraction for Web Mining," and "Deploying Approaches for Pattern Refinement in Text Mining," In these papers a technique of closed sequential patterns is used in text mining. It contains the concept of closed patterns in text mining. It improves the performance of text mining. Pattern taxonomy model is developed to improve the effectiveness. It uses closed patterns in text mining effectively. Term-based methods and pattern based methods is used to improve the performance of information filtering.

In paper **[9]** we investigate how we can assist users in the tagging phase. The contribution of our research is twofold. We analyse a representative snapshot of Flickr and present the results by means of a tag characterisation focusing on how users tags photos and what information is contained in the tagging. Based on this analysis, we present and evaluate tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. The results of the empirical evaluation show that we can effectively recommend relevant tags for a variety of photos with different levels of exhaustiveness of original tagging.

In this paper **[8]** Social tags are user-generated keywords associated with some resource on the Web. In the case of music, social tags have become an important component of "Web2.0" recommender systems, allowing users to generate playlists based on use-dependent terms such as chill or jogging that have been applied to particular songs. In this paper, we propose a method for predicting these social tags directly from MP3 files. Using a set of boosted classifiers, we map audio features onto social tags collected from the Web. The resulting automatic tags (or auto tags) furnish information about music that is otherwise untagged or poorly tagged, allowing for insertion of previously unheard music into a social recommender. This avoids the "cold-start problem" common in such systems. Auto tags can also be used to smooth the tag space from which similarities and recommendations are made by providing a set of comparable baseline tags for all tracks in a recommender system.

Manual Annotation

The most basic annotation tools allow users to manually create annotations. They have a great deal in common with purely textual annotation tools but provide some support for ontologies. There are several such programs which produce Annotea RDF mark-up. For example, the W3C Web browser and editor Amaya (Quint & Vatton 1997) can mark-up Web documents in XML or HTML. The user can make annotations in the same tool they use for browsing and for editing text, making Amaya a good example of a single point of access environment. It has facilities for manual annotation of web pages but does not contain any features to support automatic annotation. The Annozilla3 browser aims to make all Amaya annotations readable in the Mozilla browser and to shadow Amaya developments. Teknowledge4 produces a similar plug in for Internet Explorer.

Victoria Uren, Knowledge Media Institute, the Open University

They proposed a paper titled "Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art" in this paper they have examine semantic annotation, identify a number of requirements, and review the current generation of semantic annotation systems. This analysis shows that, while there is still some way to go before semantic annotation tools will be able to address fully all the knowledge management needs, research in the area is active and making good progress.

III. RELATED WORK

Existing work is done in this work format where the annotation scheme is being improved by CADS technique; here is the flow how the existing flow is working.

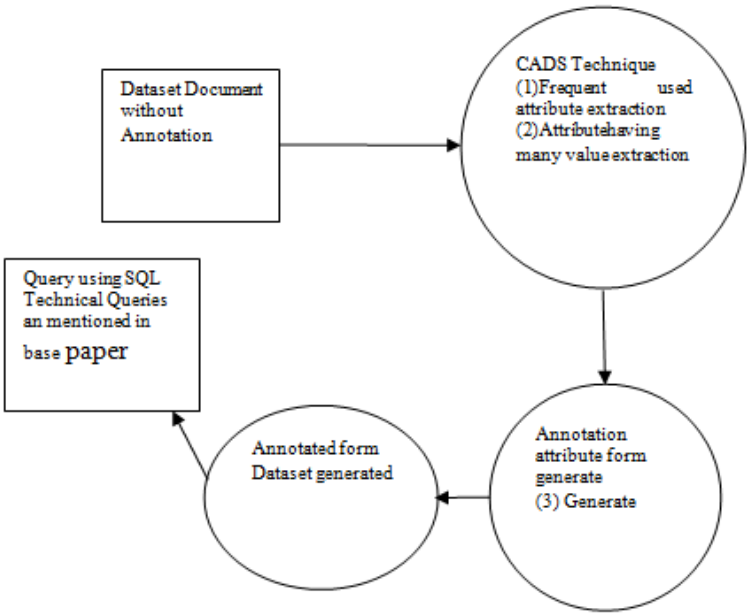


Fig-Existing System Flow

IV. ISSUES WITH PREVIOUS TECHNIQUE

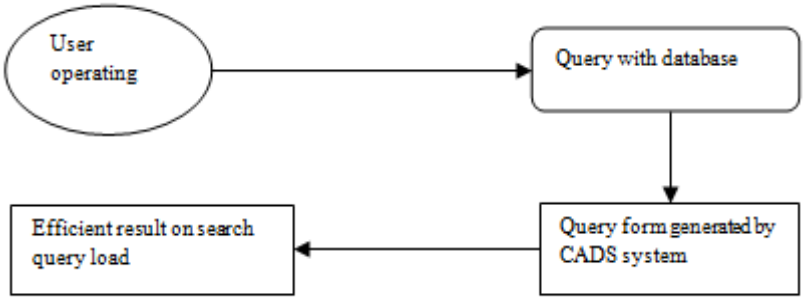
Problem with the existing work is while the efficient technique for annotation is available but still the data can only be access by the technical expertise but in the situation on demand where the possibility for the user to get the idea to access the structured data in need to avail for non-technical candidates who are often on demand requirement of these data of related to disaster or any sort of result which they often need to enquire on phone or via some short info.

The following issues have been arising with the existing technique:

- No efficient query algorithm is proposed yet.
- The available query technique is technical scope specific. Result can only be obtained based on the accurate query.
- No knowledge of data type without any prior study or knowledge about the accessed data on which user is working.

V. PROPOSED WORK

A Enhancement to the CADS where the efficient querying form also make generate where the querying to the structured dataset can be used by the normally people who might interested to access the data which is available for them, the data might be querying in the manner so that can be used by various research fellows and by different department of related data, we further can compare different querying technique by other approaches and experimentally can perform our efficient work over other work, because here we observed that number of work has been done for the structuring of dataset but still accessing them is technical task which is not for non-technical people, so here further we can extend the technique to work upon querying the dataset and also we can perform other querying technique and can compute precision, recall, query value, Computation time and computation cost.



We can definitely get good results in all the aspects as the computation time also will get less as less recourse need to provide while the data need to publish in public.

efficiency for every technique is often calculate using recall and precision and use to observe the results by different image technique while querying the dataset which we retrieved after annotation operation.

For query q,

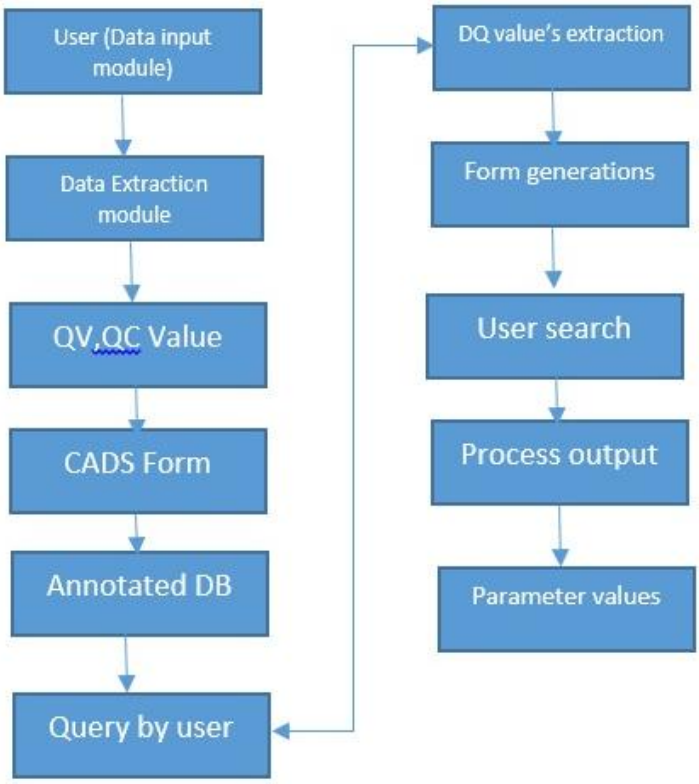
A (q) =A set of images in dataset

B (q) =A set of images relevant to query

Precision = A (q) ^B (q)/A (q)

Recall=A (q) ^B (q)/B (q)

So precision is represents the ratio of the number of images relevant to the query q among retrieved images to the number of retrieved images & Recall is the ratio of the number of images relevant to the query among retrieved images to the number of images relevant to the query in a DB, so all the evaluation process of query the dataset are always done using these two calculations. These things we can perform with different retrieval technique on CADS



Flow chart

ALGORITHM USED:-

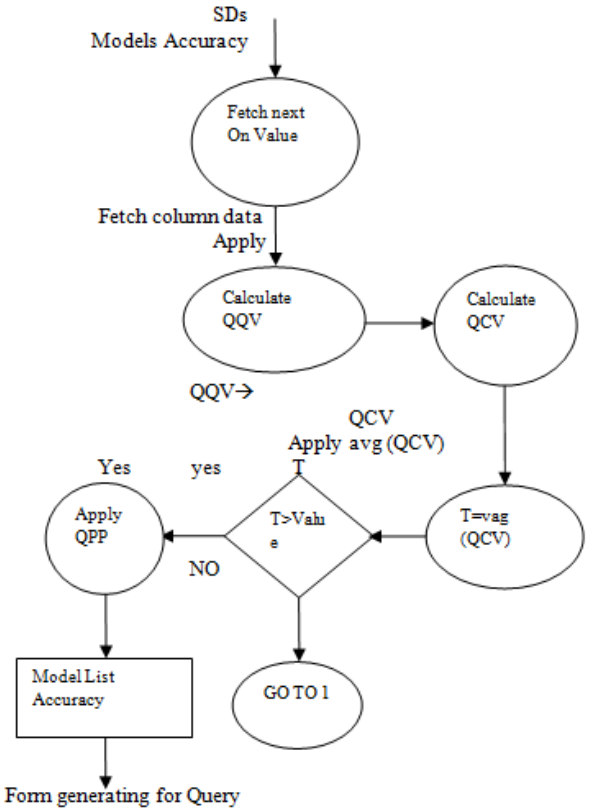
Input: SD's, Models, Accuracy

Output: Query output & QCADS Form

Steps:

1 Retrieve next on from column 1-n.
2 Get the column and data value for each column ci Calculate QQV QCV.
3 Calculate QCV.
4 Constant or Calculate T = avg (QCV). Where QCV is the maximum possible extraction of the all unseen values.
5 Sequentially process data
   i. T> value
   ii. R put value
6 Apply QPP process to get model list and adaptive accuracy which is provided in input.
7 Apply QCV
   a. Else
   b. Go to Step 1

DATA FLOW DIAGRAM:



VI. CONCLUSION

As per the discussion and the work we have discussed here we have got to know about the work which is already done in the field of annotation and accessing the dataset which is keep structured with the help of different annotation based algorithm, we have also discussed CADS algorithm which is proposed in our base paper and also we have mentioned the limitation of the existing & available algorithm, so here upon discussion we are proposing new algorithm which is efficient and going to work in the field of accessing annotated dataset.

REFERENCES

[1] S.R. Jeffery, M.J. Franklin, and A.Y. Halevy: proposed a paper "Pay-as-You-Go User Feedback for Data space Systems,"

[2] K. Saleem, S. Luis, Y. Deng, S.-C. Chen, V. Hristidis, and T. Li: proposed a paper "Towards a Business Continuity Information Network for Rapid Disaster Recovery.

[3] J. M. Ponte and W.B. Croft: proposed a paper "A Language Modeling Approach to Information Retrieval".

[4] R. T. Clemens and R.L. Winkler: proposed a paper "Unanimity and Compromise among Probability Forecasters.

[5] G. Tsoumada's and I. Vlahos's: propose a paper "Random Label sets: An Ensemble Method for Multilevel Classification.

[6] P. Heymann, D. Ramage, and H. Garcia-Molina: proposed a paper "Social Tag Prediction".

[7] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C.L. Giles: proposed a paper "Real-Time Automatic Tag Recommendation".

[8] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green: proposed a paper "Automatic Generation of Social Tags for Music Recommendation.

[9] B. Sigurbjornsson and R. van Zwol: proposed a paper "Flickr Tag Recommendation Based on Collective Knowledge".

[10] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[11] Allwein, E. L., Schapire, R. E., & Singer, Y. (2000). Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers, JMLR 1:113–141

[12] Kisilevich S., Rokach L, Y Elovici, B Shapira (2010), Efficient multidimensional suppression for kanonymity, IEEE Transactions on Knowledge and Data Engineering, 22(3):334-347

[13] Ben-David, S. and Cesabianchi, N. and Haussler, D. and Long, P.M. (1995), Characterizations of Learnability for Classes of (0... n)-Valued Functions.

