International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 04 | July-2015 www.irjet.net p-ISSN: 2395-0072

DQF for Text Analysis

Uzma Kokalgave[1], Prof. Soumitra Das[2]

[1] *Student, Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon Pune, India*
[2] *Professor, Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon Pune, India*

Abstract - ***Today's world has a high importance of data or information which has to be maintained and managed very carefully. Data storage and its access has become a challenge for us. If data is in structured format such as relational data it is easy to manage and access but if it is in unstructured format then challenges are more. While accessing this data user may write ad-hoc queries on it. To manage ad-hoc queries, predefined query forms are failing. Due to which Dynamic Query Forms (DQF) can be introduced to accomplish this task. We are proposing a system in which end user can write whatever kind of queries to analyze unstructured data that they are not aware of, to get desired queries results. In this system user can first select the input file which contains unstructured data such as sentiment data or invoice information in .csv or pdf files. Data from those files will be used for analysis and dynamic query forms will help to generate query forms, in which end user themselves will create the query form and will execute the queries. DQF provides iterative approach where user will enrich the query from until they get desired output of their queries. This system also allows users to rate the query form so that it will be ranked by system using ranking metrics such as precision and recall. This can be done by calculating F-measure. As DQF is a web application it is highly expected to give quick responses for the queries. As a future work we may use this system for image and videos.***

Key Words: *Unstructured Data, Text Analysis, Dynamic Query Form, Query Ranking and enrichment.*

1. INTRODUCTION

In most of the situation the while interacting with database we have to come across certain ad-hoc queries which are not pre-written or cannot be judged at design time. Here we require dynamic query form generation so that user can design a query form first with the concept of the query he wants to execute and then he can execute the query to get desired output. The main idea while generating dynamic query form we experience when the data to be handled is unstructured in nature. Unstructured data such as pictures, images, videos, some text files which have been stored data such as invoice info, sentiment data etc. To write ad hoc queries on such data, we can allow users to generate dynamic query form. In this work we are proposing the same to get desired output.

In DQF for Text Analysis there are following main component:

1. Input files which has unstructured data such as sentiment data.
2. Certain set of form component such as *text box, button, label, radiobutton, checkbox* etc.
3. Dynamic query form will be generated to write and execute the queries to get the desired result.
4. Ranking form for ranking a query form once user gets desired query result.

Outcomes

1. Generating dynamic query form.
2. After getting desired query result ranking of query form.
3. For writing ad hoc queries for such unstructured data DQF will be very useful which will avoid a huge set of predefined queries on the server.

Such type of system can be used in Crime Investigation Department, Meteorological or oceanography department, Mining sentiment data or opinion, mining satellite images such as scientific images etc. Previously there are various approaches given for generating DQF for relational and non-relational data. [1] [2] But in our approach we are proposing analysis of data first and then create a DQF for writing desired queries on that data. DQF also supports ranking of a query form. To examine the effectiveness of a query form for getting desired output F- measures are used. Two ranking metrics precision and recall will be used to rank the query form. [3]

2. EXISTING SYSTEM

Generating Dynamic Query Form for databases [1] adds an advantage to the end users as well as to the system. In most of the applications we found the Static query form generation can lead to overhead on system and server to maintain large number of static query forms. [4] But due to Dynamic Query Forms, server need not to maintain the large set of query forms. On the top of it DQF provides a direct interactions with the end users even though they are not aware about databases. In some work DQF has been used for relational data [1], for non-relational data

[2] and for semi structured data also [5]. Most of the technologies support customized query forms such as Microsoft access database cold Fusion [11] and Easy Query [10]. but the problem with these systems is that they are created for Professional developers and not at all for end users. We found some work done on securing DQF using CAPTCHA in dynamic query form execution [3]. Some other work has been done on query form improvement. [6] There are some surveys have been made to how one can optimize search using dynamic query forms [7]

Table -1: Name of the Table

Paper Reviewed	Parameters				
		Database Used			
	Dynamic Query form	Relational Data	Non relational Data	Unstructured Data	Ranking of Query Results
Dynamic Query Form for Database Queries	Yes	Yes	No	No	Yes
Dynamic Query form for Non relational Database	Yes	No	Yes	No	Yes
Randomized query formulation for Database Queries	Yes	No	Yes	No	Yes
DQF for Text Analysis	Yes	No	No	Yes	Yes

There is a technique called Query refinement used by most information retrieval systems [8] [9]. In this system, new terms related to the query are being recommended or the terms according to the navigation path of the user in the search engine get modified. But for the query form used for database, there are database queries not individual terms to be altered.

3. PROPOSED SYSTEM

3.1 Our Approach

In Our Proposed system for DQF we will be using unstructured data such as sentiment data from social networking sites as an input. DQF works in two phases as Query execution and Query Enrichment [1]. In Query Execution user will select the input file from local machine and then it will be used for analysis of the data, depending on users choice data will be analyzed and then it will be used in DQF for writing queries on it. In this case user first creates the query form and writes query on it and submit the query .If he satisfies with output of it then he will exit from the DQF or else second phase will be executed for Query Enrichment. In this phase user will modify the query form depending on the required query result.

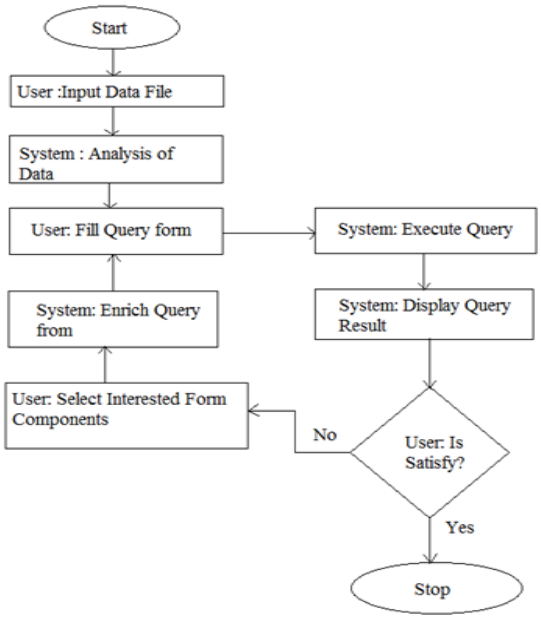

Fig -1: Flow diagram for DQF

Mathematical Model

S = {Data, DQF, Enrich, Rank}
Data = {Sentiment Data}
Rank = {Outstanding, Good, Average, Below Average, Not Good}
FC = {Labels, Text boxes, Radio buttons, check boxes, buttons}
DQF = {FC}

U: $f(S)$
$f(S) \rightarrow QR$

Where,
S: System for DQF for Text Analysis
Data: Data files used in DQF
DQF: Dynamic Query Form
Rank: Ranking DQF
FC: Form Components
U: User
QR: Query Result

f_1 (Data) → Input Data File
f_2 (DQF) → Generating Dynamic Query Form
f_3 (Rank) → Ranking DQF
f_4 (Enrich) → Enriching (Modifying) Query Form

Flow of Model

1. *Start*
2. *Login or Register*
3. *f_1 (Data)*
4. *f_2 (DQF)*
5. *if User Satisfies then f_3 (Rank) else f_4(Enrich) and f_3 (Rank)*
6. *Exit*

3.2 Ranking score estimation and ranking metrics

In our work we are allowing user to rank the query form. Based on ranking metrics such as recall and precision query form will be ranked and it will be kept for future purpose. For evaluating query results there are two measures available such as precision and recall. Based on different inputs provided to query forms can give different output in query results. To achieve expected query result we will be using expected recall and expected precision. Expected precision is the proportion of the query results which is interested by current user and expected recall is the proportion of users expected interested data instances which are returned by current query form. For ranking score estimation two components will be ranked one is projection and another is selection [1]. In projection components ranking, entities (Tables) and their respective attributes (Columns) will be ranked. Here attribute with maximum F-score will be selected. And in raking selection form components first important attribute will be selected and ranked, here the F-score would be computed incrementally on desired attributes.

4. SYSTEM IMPLEMENTATION AND EXPERIMENTAL SETUP

Dynamic query form is a web based application developed using ASP.NET using C# language. All execution of the system is supported using asp.net only. For data analysis we have used R Analytics features. Instead applying separate analytics on data, we found a package for performing same operations using R in R.NET [12]. All experiments are running on a machine with Intel ® Core ™ i5-3340M CPU @ 2.70GHz, 2.40GB main memory and running on Windows 7 Enterprise operating system (32-bit).

Datasets : We have chosen some sample database (.CSV files) for performing experiment such as Hospital revised flat files from [13], Online News Popularity from [14], Some sentiment data from twitter such as global warming tweets, airlines sentiments captured during their peak time.

Results:

The application runs in following way, User selects the input file and selects the type of analysis to be performed on the data. Refer the below figure for first screen shot.

Fig -2: Web page for selecting input file and type of analysis to be done on selected type of data.

Once data captured by the system, it helps used to generate query form with desired form components selected by user. And proceed for generating queries and results of it. Once query will be executed with desired result after certain iterations. It will be ranked by system based on what rating user has chosen for it.

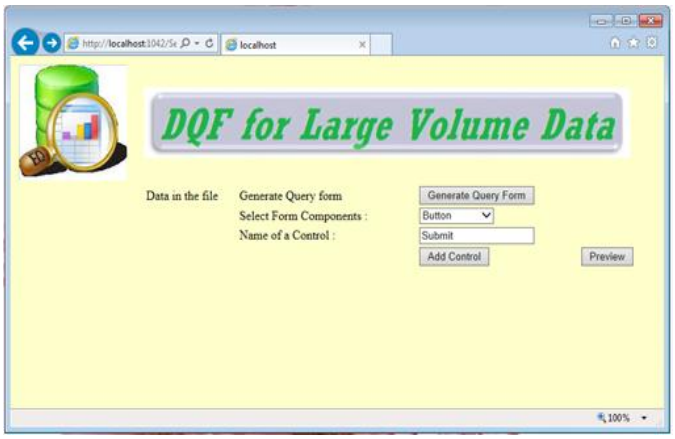

Fig -3: Web page for generating DQF

Estimation of query results can be calculated with expected recall, expected precision and F-score using following formulae.

Precision is % of selected items that are correct so formula for precision is

$$\text{Precision} = TP / (TP + FP)$$

And recall is % of correct items that are selected, so formula for it is

$$\text{Recall} = TP / (TP + FN)$$

Where, refer Table-2.

Table -2: Parameter for calculating Precision and Recall

	Correct values	Not Correct values
Selected values	TP	FP
Not Selected values	FN	TN
	Correct values	Not Correct values

And finally F measure for estimating correctness of a query form we have written a formula

F-measure = 2 x ((precision x recall) / (precision + recall))

5. CONCLUSIONS AND FUTURE SCOPE

DQF is the system in which user has flexibility to choose the unstructured sentiment data file for analysis and once analysis done on the data dynamically user can select the form component and design a query form at run time and executes the query on analyzed data to get desired output. Such system can be effectively used in areas like crime investigation, sentiment data mining etc. As a future scope we can also use it for image and video analysis and query execution.

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude and humble thanks to my guide for his expert guidance and focused direction on publishing this work. I would also like to thanks publisher and the team who are directly or indirectly helped me to accomplish this task. I want to give sincere thanks to my college for providing me setup for the completing my project.

REFERENCES

[1] Liang Tang, Tao Li, Yexi Jiang, Zhiyuan Chen, *Dynamic Query Forms for Database Queries*, IEEE Transactions on Knowledge and Data Engineering, 19 April 2013.

[2] S.BhaskaraNaik, B.VijayaBhaskar Reddy, *Dynamic Query Forms for Non-Relational Database Queries*, International Journal of Engineering And Computer Science ISSN: 2319-7242 Volume - 3 Issue -8 August, 2014 Page No. 7415-7419

[3] Gopi Krishna Lakkasani, Balakrishna Nayudori *Random Query Formulation for Database Queries* , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 8, August 2014, ISSN: 2277 128X

[4] M. Jayapandian and H. V. Jagadish. *Automated creation of a forms-based database query interface*. In Proceedings of the VLDB Endowment, pages 695–709, August 2008.

[5] VISHNU R , SWAPNA HARI , *DYNAMIC QUERY FORMS WITH NoSQL* International Journal of Research in Engineering & Technology (IMPACT: IJRET) ISSN(E): 2321 8843; ISSN(P): 2347-4599 Vol. 2, Issue 7, Jul 2014, 157-162

[6] K. Chen, H. Chen, N. Conway, J. M. Hellerstein, and T. S. Parikh. Usher: *Improving data quality with dynamic forms.* In Proceedings of ICDE conference, pages 321–332, Long Beach, California, USA, March 2010.

[7] Prajkta Dagade, Mansi Bhonsle, *Espionage on Search Optimization using Dynamic Query Form* International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014, ISSN 2091-2730.

[8] W. B. Frakes and R. A. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, 1992.

[9] Meenu Joy Bhruguram T M, *Dynamic Query Form with query Refinement and Database encryption* , IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep Oct. 2014), PP 154-159

[10] EasyQuery. http://devtools.korzh.com/eq/dotnet/.

[11] ColdFusion. http://www.adobe.com/products/coldfusion/.

[12] R.NET package: https://rdotnet.codeplex.com/

[13] Medical data: http://data.medicare.gov

[14] UCI Machine learning Sample databases : http://archive.ics.uci.edu/ml/

BIOGRAPHIES

Uzma K. has completed Bachelor degree in Information Technology from Swami Ramanand Teerth Marathwada University Nanded and Apprearing Master degree inComputer Engineering from Pune University.

S. Das received his Bachelor degree in Computer Engineering from North Maharashtra University, Jalgaon, Maharashtra, India and Master degree in Computer Engineering from University of Pune, Pune, Maharashtra, India. Currently, he is PhD researcher at Sathyabama University, Chennai, India. His research interest includes Computer Networks,Wireless Sensor Networks, etc. He is a members of IEEE, CSI, LMISTE, IACSIT and IAENG. He is also an active reviewer of various conferences and journals.

