Prediction of Hard Queries using Keyword Classifier over Databases

Aakana Naresh Babu[#] **S V Suryanarayana***

[#]Student , M.Tech (CSE) , GVIT. Tundurru P.O, Bhimavaram
**Associate Professor (Ph.D) , Dept. of CSE , GVIT . Tundurru P.O, Bhimavaram*

---***---

Abstract*:* In modern days, there is huge collection of data in day by day. So maintaining these petabytes of data is one type of challenge and getting the results from those enormous data is challenging task to any computer professionals. Finding a particular keyword from those datasets and arranging the results is misnomer. We concentrate on data retrieval part with user keyword combinations along with ranking. Keyword query search in databases will give efficient results to the users. Searching over database is based on the queries. Sometimes we may not get the effective results to the users because it gives unwanted data and query couldn't predict the exact keyword related answers. In this paper, we propose a novel frame work for ranking the query results and predict the keyword combinations with synonyms using classification model. Classification model is used to describe the discrete variables of particular tuples in the given database. Ranking robustness principle is used for both structured and unstructured data. We built a classifier called keyword classifier which describes some accuracy rules over set of keywords imposed on query

Keywords : Classifier , dataset , Structured Robustness, attributes

INTRODUCTION

Keyword Query search is really typical tasks to search engines. Data can be reside in either structured or unstructured format. The main aim of this work is to retrieve the user interested data from the data bases for hard queries. Queried results are ranked so that the required data is shown at the top of the search. Data sets contains entities and entities contains set of attributes. if we ask the query as the required attribute the answers are very predictable . other wise it will search all related documents of particular attribute entities. Datasets can be present in either xml formats or relatioinal data format. INEX workshops data set having data in the form of xml (structured data) it is collection of structured data. for example KQI must find the desired attribute associated with each keyword in the query. For example query q god father in IMDB database ,it doesn't specify whether user asks for movie or director or distributer of particular attribute it collects all relevant information of the attribute from data sets and produce to the users. Results proven that even in structured data it is very hard to predict subset of queries with ranking. INEX and semantic search gives almost equal equal results. If we consider a keyword present in all documents we couldn't provide effective ranking to the documents. In this scenario the user interestingness was measured before search the data with combination of keywords. Predicting hard queries from unstructured documents can be done in two ways.

1. Pre retrieval and
2. Post retrieval methods

In pre retrieval, predicting hard queries is done without computing results, internally it uses statistical analysis
Post retrieval methods utilize results of a query to predict its difficulty .it has three approaches

1. Clarity score based
2. Ranking score based
3. Ranking robustness

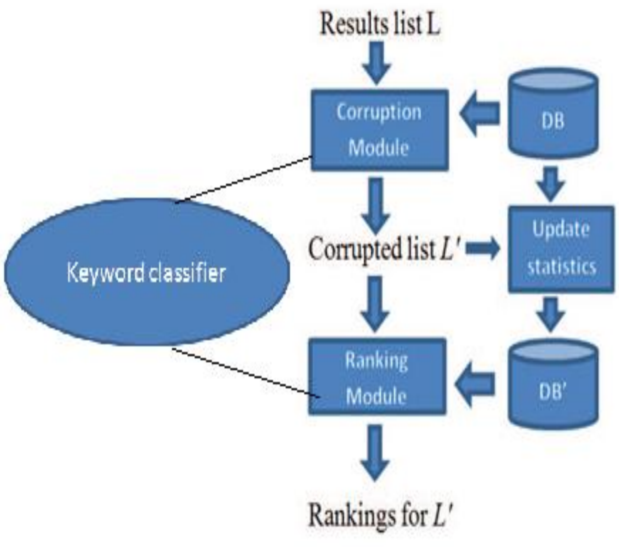
Clarity score based: in this approach only few documents are retrieved based on query difficulty so it is very easy predict the results Ranking score based: they measure the degree of difficulty of a query by computing difference between weighted entropy of top ranking result scores and weighted entropy of documents

Robustness based: there is a negative correlation between the difficulty of a query and its ranking robustness in the presence of noise in the data

2. Estimation techniques for ranking the query

Query results are shown in some order called as top K Results in the database. A structured data set contains large amount of information stored in structured format say XML like in Data centric workshops (INEX). We usually corrupt the top K results and re ranked the results over corrupted data base

3. Architecture



In this model we extend S-R Algorithim by adding a classifier. The classifier is built from the training set made of data base tuples along with the given keyword. a keyword classifier was added to the existing ranked algorithm to easily predict the hard queries using some predefined classification rules for given keyword list. If the results having more corrupted data then we use this keyword classifier efficiently. SR Algorithm generates noise while query processing. Results says that SR algorithm works slow but our modified model will improve the accuracy or ranked results.

Keyword Classifier

Keyword classifier is a classifier model to generate set of frequent rules for ranking the results. Users search may contain two or more keywords in the same query, based on the keyword it built a model by considering the synonyms of particular keywords and combination of keywords. As we know the classification is for discrete valued attributes. It searches the keywords from documents, datasets and entities of particular attributes. Structured robustness algorithm generates the top K ranked results based on this model.

Keyword classifier algorithm in query evaluation

Step 1: Select input as tokens T

Where T = { t1, t2, t3,.. }

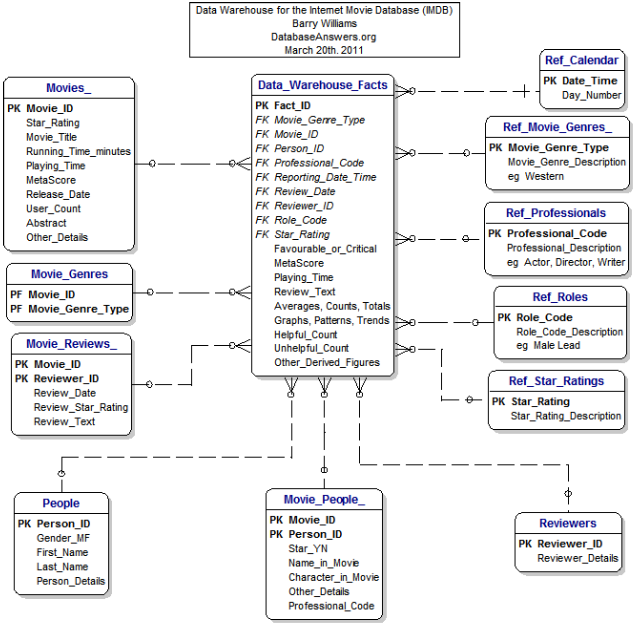
Step 2: generate most frequent hits for combination of tokens K

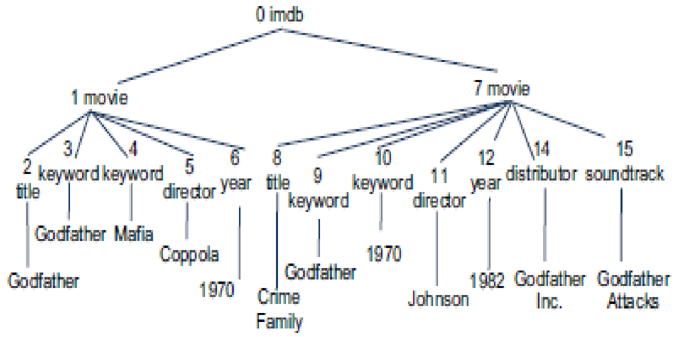
Step 3: Build a classifier called keyword classifier along with the count

Step 4: retrieve data based on the K

Step 5: rank the results using SR Algorithm

Consider movie database IMDB,





In this above data the keyword godfather is considered with different combinations like movie, director, and actor and generates the frequency count for each combination then the results will be projected. By using keyword classifier it is very easy to understand and the results are almost similar without using any approximation algorithms. The key is the usage of xml while using xml documents. We can easily navigate tag attributes from one object to another. The more key terms matched will give the best results in keyword classifier approach

S R Algorithm

Structured Robustness algorithm generates the noise over database during query processing. It loops over all attributes in corrupted list to check whether it is corrupted or not. One entity may have hundreds of attribute values. So it takes considerably some more time to work SR Algorithm has to re rank the top k entities for corruption. SR Algorithm takes long time to loop that re ranks the corrupted result. Along with keyword classifier it assigns a frequent count of each keyword in the documents to rank the results. It will generate better results without using any approximation algorithms for improving structured robustness.

Future work

In future, we may get data with ranked attributes list selection so that we can retrieve data exactly what we need without standard inputting. This technique may helpful in smart device computing.

Conclusion

In this paper, we propose an efficient SR Algorithm without finding approximations. We built a novel framework classifier called keyword classifier. Keyword classifier works on combination of keywords to built a model and generate the corrupted results of top K Ranked queries efficiently. Using of keyword classifier will generate the frequent occurring top K Ranked queries over corrupted data bases. Consider god father as keyword in IMDB Database. It generates set of rules over movies directors and actors and rank the results in appropriate manner. This model is applicable to both structured data and normal text documents.

References:

[1] Shiwen Cheng; Termehchy, A.; Hristidis, V., "Efficient Prediction of Difficult Keyword Queries over Databases," in Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.6, pp.1507-1520, June 2014

[2]Y Lou , X Lin , W Wang and X Zhou , “SPARK” Top –k keyword query in relational databases

[3] V Hristidis , L Gravano and Y Papakonstantinou “ Efficient IR Style keyword search over relational database “

[4]J Kim X Xue and B Craft “ A Probabilistic retrieval model for semistructured data” ECIR, Tolouse , France

[5]A Trontman and Q Wang “ Overview of INEX 2010 data centric track “ Workshop INEX 2010 , Netherlands

[6]N Sarkas S Paparizos and P Tsaparas “ Structured Annotations of web queries “ ACM SIGMOD , USA

[7] S C Townsend Y Zhou and B Croft “ Predicting Query Performance “ , Finland

[8] J A Aslam and V Pavlu “ Query Hardness Estimation using Jensen- Shannon divergence among multiple scoring functions , ECIR , Rome Italy

[8]O Kurland A Shtok , D Carmel and S Hummel “ A Unified framework for post retrieval query performance predictions “ , ICTIR, ITALY 2011