International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 3366–3372

Elimination of Duplicate Records from Multiple Resources by Applying Normalization

Mohini Markad[1], Renuka Shaikh[2], Shivani Chaudhari[3], Mayuri Patil[4]

JSPM's Rajarshi Shahu College of Engineering, Pune, India

[1]mohinimarkad1996@gmail.com, [2]renushaikh24@gmail.com,
[3]shivanirchaudhari1@gmail.com, [4]patilmayuri472@gmail.com,

Abstract

Data duplication is a big issue in various organizations. Data redundancy is the major cause of memory wastage and time wastage. Most of the times data is remaining in the system even if it is of no longer use. At large level it becomes impossible to identify such data and organizations end up spending huge amount of money and time. More money is spent on buying clouds for information storage instead of clearing unused previous stored data. Normalization helps a lot in this problem. The rate of learning can be increased in normalized record as compare to non-normalized data. Also if we add some time server for the information then we do not even need to remember. At certain time files will be deleted automatically. Memory and Time can be saved. Then Here we mine big data using Keyword search algorithm for getting duplicate records from the database. We created a system for detecting the redundant data. The invented system includes algorithms to get duplicate records and save only normalized records. Here we use time server that are used for time and storage utilization. When time limit exceeds then that file automatically delete from database. Here we provide more security by applying Encryption on the files and detect duplicate records using file signature or tag. Here experimental result shows our system are more efficient than existing systems

***Keywords*:** *Record normalization, searching, De-duplication, time server*

1. Introduction

Integration system as we scale got to automatic matches record from the different sources that ask an equivalent real-world entity to find truth matching the record among them and switch this sets of record into a typical record for the consumptions of the user or other application. There is an outsized body of the labors on the records matching problems and wherefores, the truths discovery problems. The records matching problem is mentioned in data duplicate records detection system, records linkages, objects identifications, entities resolutions, reduplications and therefore the truth of discovery problems called truth finding and fact finding key problems in the data fusion. In this paper, we assumes task of record match data and truth's discovery are performs in which the group of the true matching record.

A normalization record is various applications. For instance, within the research publications domains, although the integrated websites, like Cite seer or Google scholar, contain the records gathers from a spreads of the source using the displayed on a normalized record to users. Record normalization may be a severe problem because

ISSN: 2233-7857 IJFGCN

different online sources have represent the attribute values of an entity in several ways or maybe provide confusing data. Confusing data may occur due to incomplete or empty data, different data representations such as missing values, and even error prone data. Data normalization is an vital in many applications .For e.g., within research publications domains, integrator website, like Cite seer and Google Scholar, contain the record gathered from the spread of source using the automated extractions techniques, it must display a normalized data or record to users. Otherwise, it's unclear what is often presented to users: (I) It will present the whole group of matching record (ii) simply presents some random records from the files. From the groups to only name of ad-hoc approach. Either of those choices can because a frustrating experience for a user because in (me) the user must be sort or browse through a potentially sizable amount of duplicate record, (ii) we run the danger of presenting a data with missing or inaccurate piece of knowledge.

2. Literature Survey

"Normalization of the duplicate records from multiple sources" Dong Yangquan, C Eduard. Dracut, Member, and Wii Meng, **"Normalization of the duplicate records from multiple sources"** Senior Member of IEEE[1] from this paper, the automated extractions techniques, it must be standardization ways, from the naive , that has been use solely info got from the records and advanced methods that generates the gaggle of duplicate records before choosing the price of the associated attribute of the records. We have a bent to conducted intensive empirical studies with all the projected ways. we have a bent to point the weakness and strength of each of them and that is to be used in observe Incremental Records Linkage Gruenheid Anja Zurich ETH, Luna Xin Dong, Srivastava Divesh ,In this paper present an associate end to end frameworks which can be increments and efficient update linkage results once knowledge update arrive. Our algorithm not solely enable merge the records within the update with in existing cluster, however, conjointly enable investing new proof the updates for repairing previous linkage errors. Experiment to be performed that's results on the 3 real or artificial knowledge sets a show that our algorithms will considerably reduce linkages times while not sacrifices the quality of linkage.

"**In Online Orders of the Overlap the Data Sources, Mariam Salloum" , Dong Luna Xin, Srivastava Divesh ,Tsotras J. Vassilis** [2] the system of data integration that offer the consistent interfaces for the querying an outsize number of the autonomous and heterogeneous sources of data. Basically, the answer are returns whenever the sources are queried, therefore this answer list is updated for more answers arrive. To choose the honest ordering during which sources is queried for critical and increasing the speed which answer are returned. How, this problem has been challenged since we frequently don't have any completed or précised for the statistic of the source, like the coverage's and overlaps. It's the exacerbated with in the Big- Data Era, which is the witness of the two trend in the Web Data that obtains a full coverage of knowledge during the particular domains often it require extracted data from thousands of sources and second is the there is often an enormous variations in the overlaps between the differential data sources. In this, we presents oasis, web query that answering the System for overlapping the Sources.

ISSN: 2233-7857 IJFGCN

"**Merg the query results from local search engine for Geo-referenced object" Dasgupta Bhaskar, Beirne P. Brian ., Atassi Ali Neyestani, Badr** [3] The Emergences of various online source about the local services presents requirement more automated yet accurate data integrations techniques. Local services are geo-referenced object and may be query by their locations on a map, as an example, neighborhoods. Typical local services queries this , we address three key problems translation merging, and ranking the Most local search engines provide a hierarchical organization of cities into neighborhood and the area in one local program may correspond to sets of neighborhoods in other local search engines. integrated access to the query results returned by the local search engines, we'd like to mix the results into one list of results. Our contributions include: (1) an integration algorithms for the neighborhoods. (2) A effective business listings resolutions algorithms. (3)The ranking of the algorithms that take into considerations the user's criteria, users ratings and ranking. We created a prototype systems over local searching engine within the restaurants domains. The restaurants domains may be representative case studies for the local services. We conducted a comprehensive experimental study to Yumi gauge.

"**A prototype versions of the gauge is out there online.NADEEF/ER: Generic and Interactive Entity Resolutions", Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani[**4] Entity resolutions, the method of to identify and eventually merging record that ask equivalent real-world entities, is a crucial long-standing problem. We presents needed /Er generic or interactive entity resolutions system, which is the made as an extension over our open source generalizing data cleaning system Nadeem. Nadeem/Er provided an upscale programs interfacing the manipulating entities, which allows generic, efficient and extensible ER. during this demo, users will have the chance to experience the subsequent features (1) Easy specifications Users can be easily defines ER rule with a browser based specifications, which can then be automatically transformed functions, treated as black box by Nadeef; (2) Generality and extensibility Uses can be customizes their ER rule by refining and fine tuning the above function to realizes both of effectively and efficiently ER solution; (3) Interactivity : We also extends the prevailing Nadeef dashboards with in the summarizations and clustering's techniques to facilitating understanding problems faced by the ER processes can also be on allow users to influence resolutions decisions

"**A Sample or Clean Frameworks for the Fastest and Accurate Query Processing on Dirty Data", Wang Jiannan, Krishnan Sanjay, Michae Ken Goldberg, Tova Milo** [5] Aggregate query processing over very large datasets is often slow and susceptible to error thanks to dirty (missing, erroneous, duplicated, or corrupted) values. To deal with the speed issue, there has lately been a resurgence of interest in sampling-based approximate query processing, but this approach further reduces answer quality by introducing sampling error. In this paper, we explores an intriguing opportunities that sampling presenting, namely, that when integrates with data cleaning, sampling actually improve answer quality. Data cleaning requires either domain-specific software or human inspections. The latter is increasingly feasible with crowdsourcing but are often highly inefficient for giant dataset Our result suggest the estimated values can rapidly converge toward truth values with surprisingly few clean sample, offering significant improvement in cost over cleaning all of the info significant improvement in accuracy over cleaning the none of them info.

3. Proposed System

Here we tend to projected levels of the normalization granularities (record-level (File Level) and field-level (Block Level)) and two types of the normalization. For multi strategy approach, we tend to used result merging models impressed from Meta looking

ISSN: 2233-7857 IJFGCN

out to mix the results from variety of single ways. We tend to analyze the records and field level of normalization within the typical normalization. Within the complete normalization, we tend to centered on field values and projected algorithms for word form enlargement and keyword mining to provide abundant improved normalized field values. Here we use AES algorithm to encrypt file and store in blocks also check duplicate files as well as blocks in database. For checking blocks we use MD5 algorithm and for diving the file into blocks we refer chunking technique. Here we use time server that are used for time and storage utilization. When time limit exceeds then that file automatically delete from database and it is used to memory storage utilization.

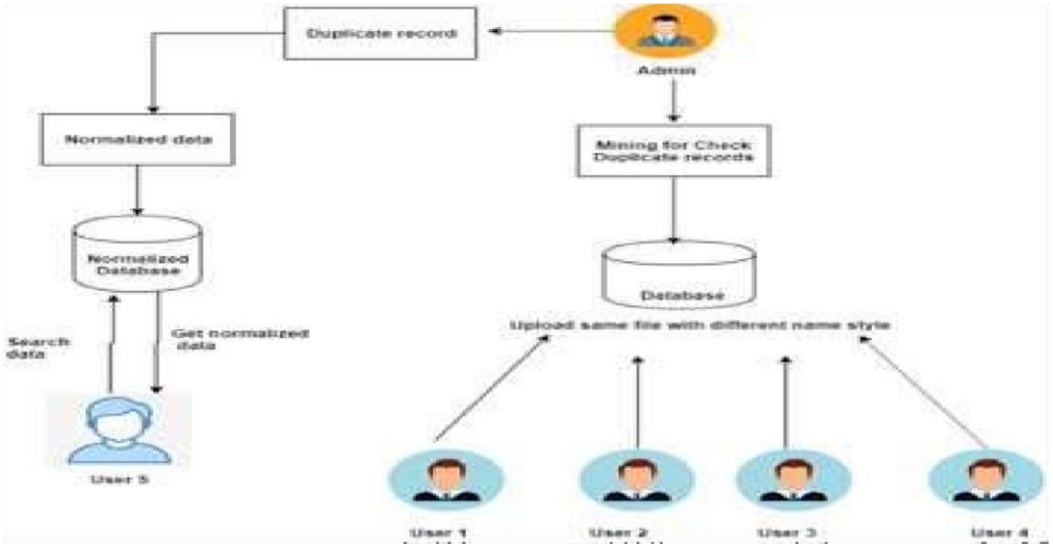

Figure 1. System Architecture

4. Algorithm

AES Algorithm:

- Derive set of round keys from the cipher key.
- Initializing the state array with the plaintext.
- Adding the initial round key to the starting state array.
- Performing he nine rounds of state manipulation.
- Performing the tenth and final round of state manipulation.

MD5: Message Digest Algorithm:

- MD5 Algorithm that was developed with motivation of the security as it take as an input of the any size and produce an outputs if a 128 bit hash values. To be consider secure MD5 should meets two requirements :

ISSN: 2233-7857 IJFGCN

- That is impossible to generate an two inputs that cannot be produce the same hash function.
- That is impossible to generate a messages having the same hash values.
- MD5 it is developed to store one way hash of a passwords and some files servers also provides
- user can compare the checksum of the downloaded file.

.

5. Math

The dataset is another factor that needs to be considered during the implementation of VQA. Proper datasets with enough images and question-answer pairs are difficult to create or obtain as it is tedious task. Standard datasets for the VQA task is MSCOCO dataset and VQA v1, v2 dataset. Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). First, use the equation editor to create the equation. Then, select the "Equation" markup style. Press the tab key and write the equation number in parentheses. To make your equations more compact, you may use the solidus (/), the esp. function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$\int_{0}^{r} F(r,\theta)\, dr d\theta = [\theta r_2/(2\theta_0)]$$

$$\theta \int_{\theta_0} \exp(-\theta\theta | z_j - z_i |) \theta J_1(\theta r_2)\, J_0(\theta r_i) d\theta .$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (T might refer to temperature, but T is the unit tesla). Per IJSTR, please refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) shows ... ." Also see The Handbook of Writing for the Mathematical Sciences, 1993. Published by the Society for Industrial and Applied Mathematics, this handbook provides some helpful information about math typography and other stylistic matters. Please note that math equations might need to be reformatted from the original submission for page layout reasons. This includes the possibility that some in-line equations will be made display equations to create better flow in a paragraph. If display equations do not fit in the two-column format, they will also be reformatted. Authors are strongly encouraged to ensure that equations fit in the given column width.

6. Conclusion

ISSN: 2233-7857 IJFGCN

We have created three levels of normalization. Those are record-level, field-level and value component level. The text file will be first checked name wise then text inside it. Here we check duplicate records by using MD5 algorithm and providing more security on file we encrypt file and store in database. For the security purpose we have created a system in which each individual's data will be encrypted so that no other person can see it and authenticity will be maintain. The experimental results shows the normalization of duplicate records, storage of unique values & effective authenticity. Also the time server which we have added performs a major role in memory saving

Table 1. Results of different approaches

NAMES	FILE UPLOAD	FILE SEARCH	F ILE DOWN LOAD
ENCRYPTION	1.8	1.7	1.7
SPLIT	1.2	1.1	1.1
DECRYPTION	1.7	1.6	1.6

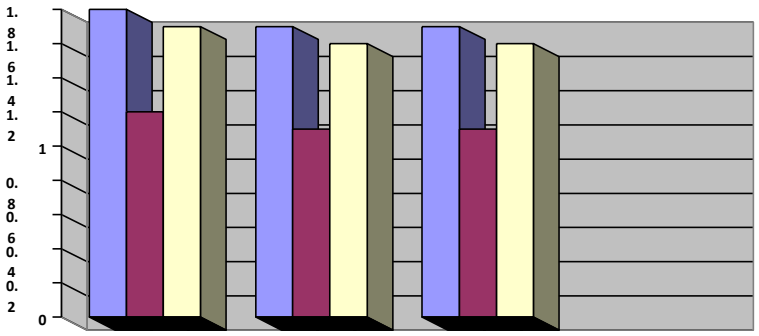

Figure 2. Comparative Analysis of Accuracies

ISSN: 2233-7857 IJFGCN

Acknowledgments

We are deeply grateful to our project guide Prof. G. D. Upadhye for the help in the field of Data Mining and its approaches.

References

1. Yongquan Dong, Eduard C.Dragut, Member, IEEE, and WeiyiMeng, Senior Member, IEEE, "Normalization of duplicate records from multiple resources", 2018, pp:1
2. M. Salloum, X. L. Dong, D. Srivastava, V. J. Tsotras, Online ordering of overlapping data sources, PVLDB 7 (3) (2013) 133–144.
3. E. C. Dragut, B. DasGupta, B. P. Beirne, A. Neyestani, B. Atassi, C. T. Yu, W. Meng, Merging query results from local search engines for georeferenced objects, TWEB 8 (4) (2014) 20:1–20:29.
4. A. Elmagarmid, I. F. Ilyas, M. Ouzzani, J.-A. Quian´e-Ruiz, N. Tang, and S. Yin. NADEEF/ER: Generic and interactive entity resolution. In SIGMOD, pages 1071–1074, 2014.
5. E. K. Rezig, E. C. Dragut, M. Ouzzani, and A. Elmagarmid. Query-time record linkage and fusion over web databases. In ICDE, 2015.
6. A. Culotta, M. Wick, R. Hall, M. Marzilli, and A. McCallum, "Canonicalization of database records using adaptive similarity measures," in SIGKDD, 2007, pp. 201–209.
7. O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: A generic approach to entity resolution," VLDBJ, vol. 18, no. 1, pp. 255–276, 2009.
8. M. L. Wick, K. Rohanimanesh, K. Schultz, and A. McCallum, "A unified approach for schema matching, reference and canonicalization," in SIGKDD, 2008, pp. 722–730.
9. L. Wang, R. Zhang, C. Sha, X. He, and A. Zhou, "A hybrid framework for product normalization in online shopping," in DASFAA, vol. 7826, 2013, pp. 370–384.
10. S. Chaturvedi and et al., "Automating pattern discovery for rule based data standardization systems," in ICDE, 2013, pp. 1231–1241.
11. E. C. Dragut, C. Yu, and W. Meng, "Meaningful labeling of integrated query interfaces," in VLDB, 2006, pp. 679–690.

ISSN: 2233-7857 IJFGCN