IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES

KARRI THANUSHA DEVI [#1], **V.SARALA** [#2]

[#1] MSC Student, Master of Computer Science,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

Abstract

Data consolidation is a challenging issue in data integration. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The promise of Big Data depends upon addressing several big data integration challenges while linking large amount of data into single file. There must be a lot of work for this data consolidation to achieve the normalized data. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this proposed application, we mainly try to normalize the records and try to find out the duplicate record attributes and then try to normalize that duplicate files while are present in database. We conducted extensive empirical studies with all the proposed methods. We indicate the weaknesses and strengths of each of them and recommend the ones to be used in practice.

1. INTRODUCTION

THE Web has evolved into a data-rich repository containing a large amount of structured content spread across millions of sources. The usefulness of Web data increases exponentially (e.g., building knowledge bases, Web-scale data analytics) when it is linked across numerous sources. Structured data on the Web resides in Web databases [1] and Web tables [2]. Web data integration is an important component of many applications collecting data from Web databases, such as Web data warehousing (e.g., Google and Bing Shopping; Google Scholar), data aggregation (e.g., product and service reviews), and meta searching [3]. Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity [4],

[5], [6], find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications.

There is a large body of work on the *record matching problem* [7] and the *truth discovery problem* [8]. The record matching problem is also referred to as duplicate record detection [9], record linkage [10], object identification [11], entity resolution [12], or deduplication [13] and the truth discovery problem is also called as truth finding [14] or fact finding [15] - a key problem in data fusion [16], [17]. In this paper, we assume that the tasks of record matching and truth discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. We call the generated record the *normalized record*. We call the problem of computing the normalized record for a group of matching records the *record normalization problem* (RNP), and it is the focus of this work. RNP is another specific interesting problem in data fusion.

PROBLEM STATEMENT

The promise of Big Data depends upon addressing several big data integration challenges while linking large amount of data into single file. There must be a lot of work for this data consolidation to achieve the normalized data. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this proposed application, we mainly try to normalize the records and try to find out the duplicate record attributes and then try to normalize that duplicate files while are present in database.

PURPOSE

In the proposed work Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data. For example, Table 1 contains four records corresponding to the same entity (publication). They are extracted from different websites. Record *Rnorm* is constructed by hand for illustration purposes.

OBJECTIVE

In proposed System, we try to implement normalization of duplicate records which are present in the DBMS,Here we try to apply record level normalization and try to remove all the duplicate records which are present in the database.In general several data owners try to upload different documents and if any document is already present in the database,this record level normalization will try to identify such duplicate cells and then identify the necessity of normalization.The main scope for designing this current application is to over come the problem which is faced in current networks. Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or (ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) we run the risk of presenting a record with missing or incorrect pieces of data.

2 . LITERATURE SURVEY

INRODUCTION

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language used for developing the tool. Once the programmers start building the tool, the programmers need lot of external support. This support obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into for developing the proposed system.

RELATED WORK

“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.

"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. EXISTING SYSTEM

The most existing methods of data normalization and detecting duplicate records is done based on the quantitative features of the individual record. In some cases these features can be easily identified and on other side some data cannot be normalized easily. In the existing system there is no accuracy for normalizing the records and finding the duplicate records and tuples. In the existing work, the system uses only Field-level normalization. There is no Integration system at Web scale which needs to automatically match records from different sources that refer to the same real-world entity.

LIMITATION OF EXISTING SYSTEM

The following are the limitations that takes place in the existing system. They are as follows:

1. There is no method to identify the duplicate records automatically.
2. In the existing work, the system uses only Field-level Normalization.
3. There is no Integration system at Web scale which needs to automatically match records from different sources that refer to the same real-world entity.
4. There is no accurate and efficient way to normalize the data from the data source.

4. PROPOSED SYSTEM

In general it is very problematic for data consolidation to achieve the normalized data. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this proposed application, we mainly try to normalize the records and try to find out the duplicate record attributes and then try to normalize that duplicate files while are present in database. We conducted extensive empirical studies with all the proposed methods. We indicate the weaknesses and strengths of each of them and recommend the ones to be used in practice.

ADVANTAGES OF THE PROPOSED SYSTEM

The following are the advantages of our proposed system.

1. The system is very fast due to identification of three levels of normalization granularity such as record, field, and value component.
2. An Exact Duplicate records detection can be done using proposed system.

3. The proposed system will do normalization of data sources in accurate manner.

It is very efficient and practical to use this proposed mechanism for normalization of duplicate records

5. IMPLEMENTATION

5.1 ADMIN MODULE

In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as

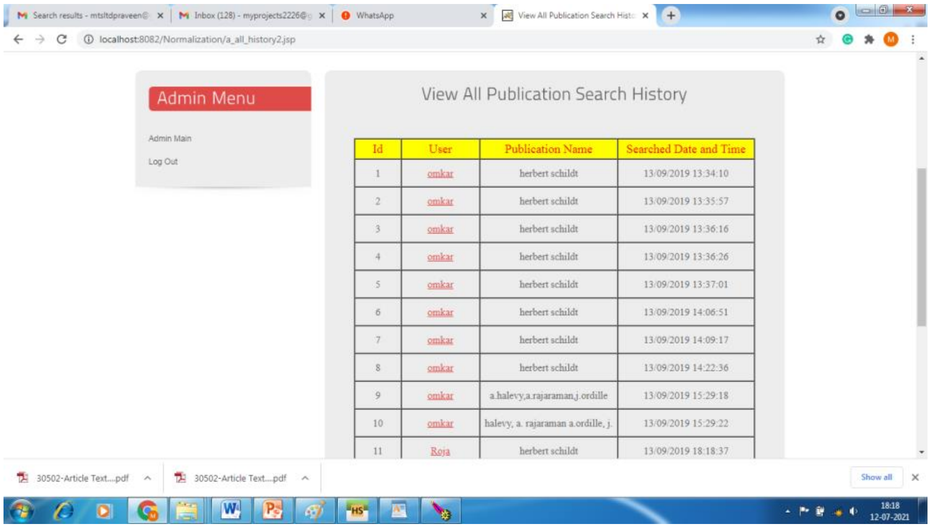
1. View All End Users and Authorize,
2. View All Uploaded Publications,
3. View All Duplicated Publication Records,
4. View All Normalized Publication Records
5. View All Uploaded Bookmarks,
6. View All Bookmark Search History,
7. View All Publication Search History,
8. View Bookmark Frequency Ranking,
9. View Publication Frequency Ranking,
10. View Rank on Bookmark in Chart,
11. View Rank on Publication in Chart.

5.2 USER MODULE

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like

1. Search Publications,
2. Search Bookmark,
3. View Bookmark Search History, and View Publication Search History.

6. OUTPUT RESULTS


Search results - mtsltdpraveen@
Inbox (128) - myprojects2226@
WhatsApp
View All Publication Search Histo
localhost:8082/Normalization/a_all_history2.jsp
Admin Menu
Admin Main
Log Out
30502-Article Text....pdf
30502-Article Text....pdf
Show all
18:18
12-07-2021

View All Publication Search History

Id	User	Publication Name	Searched Date and Time
1	omkar	herbert schildt	13/09/2019 13:34:10
2	omkar	herbert schildt	13/09/2019 13:35:57
3	omkar	herbert schildt	13/09/2019 13:36:16
4	omkar	herbert schildt	13/09/2019 13:36:26
5	omkar	herbert schildt	13/09/2019 13:37:01
6	omkar	herbert schildt	13/09/2019 14:06:51
7	omkar	herbert schildt	13/09/2019 14:09:17
8	omkar	herbert schildt	13/09/2019 14:22:36
9	omkar	a.halevy,a.rajaraman,j.ordille	13/09/2019 15:29:18
10	omkar	halevy, a. rajaraman a.ordille, j.	13/09/2019 15:29:22
11	Roja	herbert schildt	13/09/2019 18:18:37

Figure . Admin View Duplicate Publications

Admin Can see Bookmark Rank in Chart

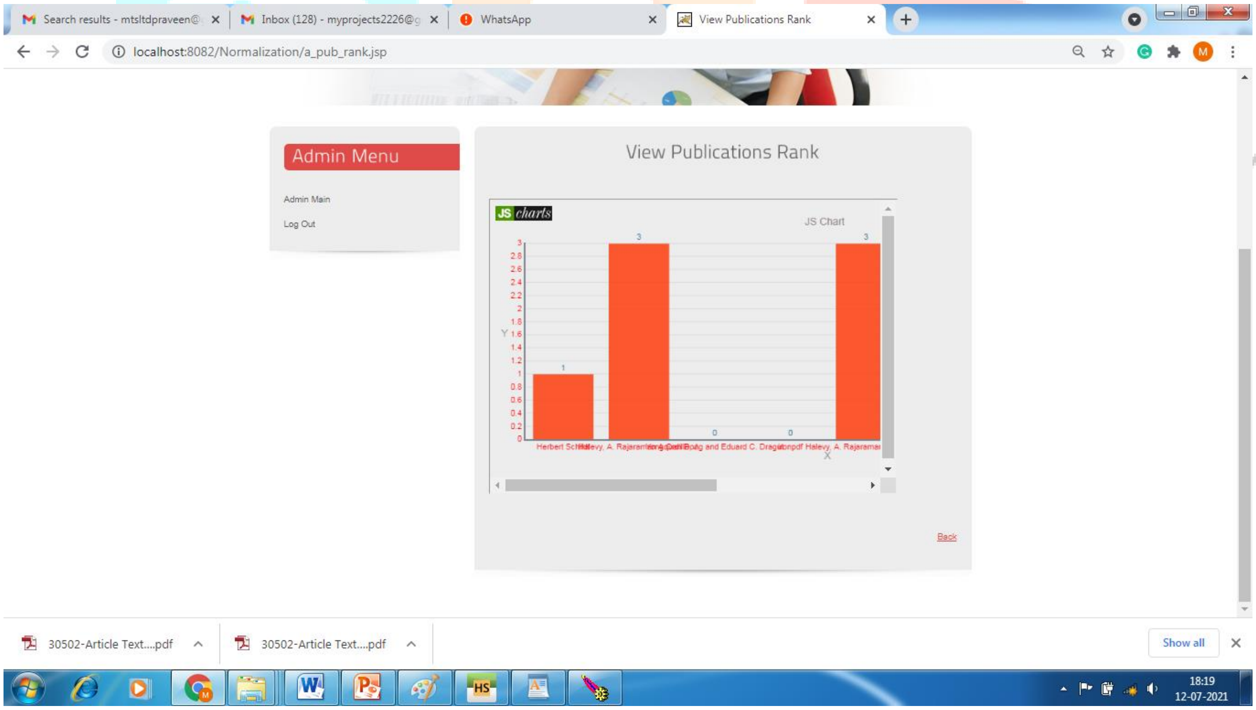
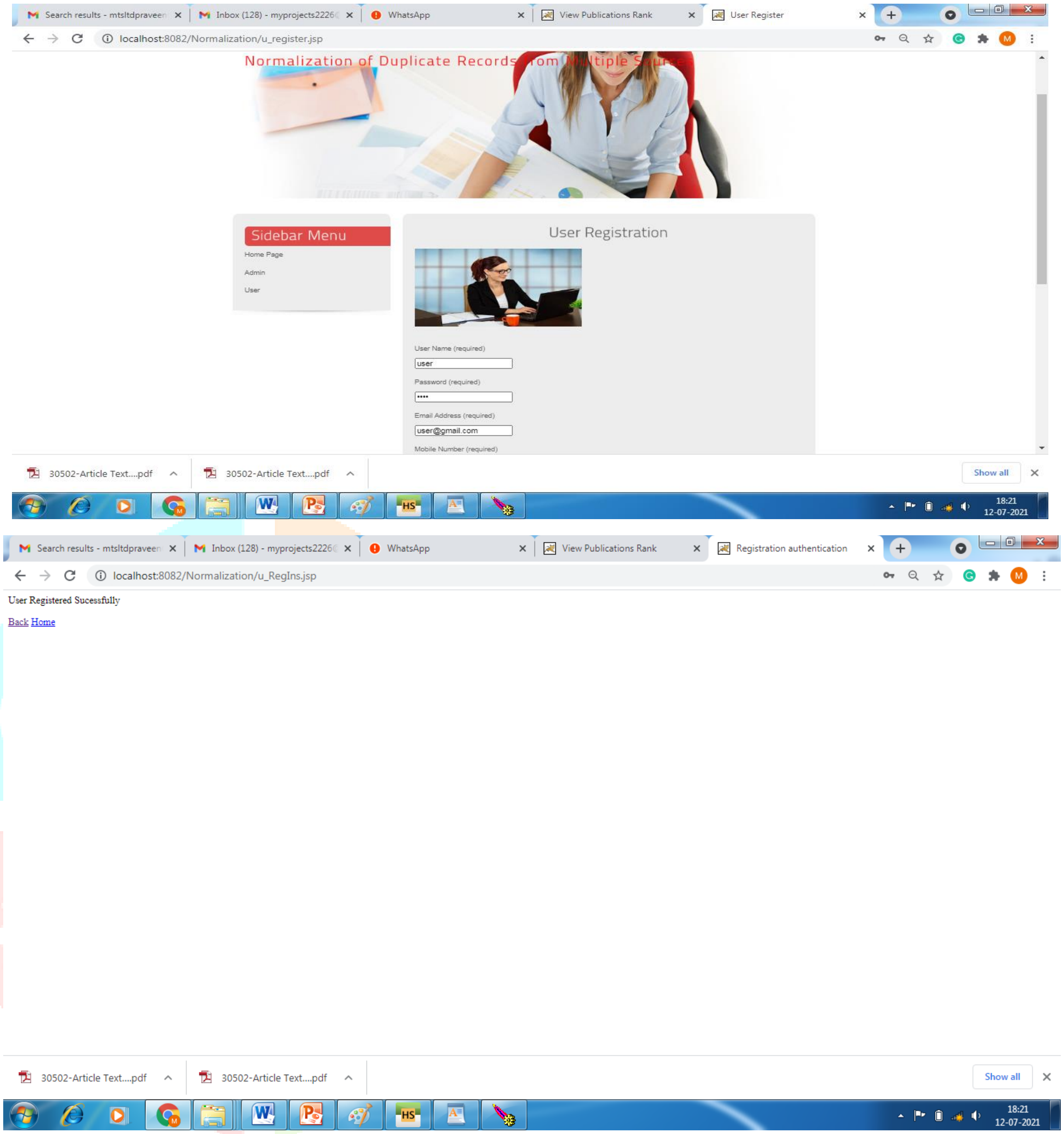

Search results - mtsltdpraveen@
Inbox (128) - myprojects2226@
WhatsApp
View Publications Rank
localhost:8082/Normalization/a_pub_rank.jsp
Admin Menu
Admin Main
Log Out
View Publications Rank
JS charts
JS Chart
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Y
1
3
0
0
3
Herbert Schildt
Halevy, A. Rajaraman, A. Ordille, J.
Hong Deng and Eduard C. Dragut
Jonpdf
Halevy, A. Rajaramar
X
Back
30502-Article Text....pdf
30502-Article Text....pdf
Show all
18:19
12-07-2021

Figure Admin View Rank in Chart Manner

User Registration


localhost:8082/Normalization/u_register.jsp
Normalization of Duplicate Records from Multiple Sources
Sidebar Menu
Home Page
Admin
User
localhost:8082/Normalization/u_RegIns.jsp

User Registration

User Name (required)
user

Password (required)
••••

Email Address (required)
user@gmail.com

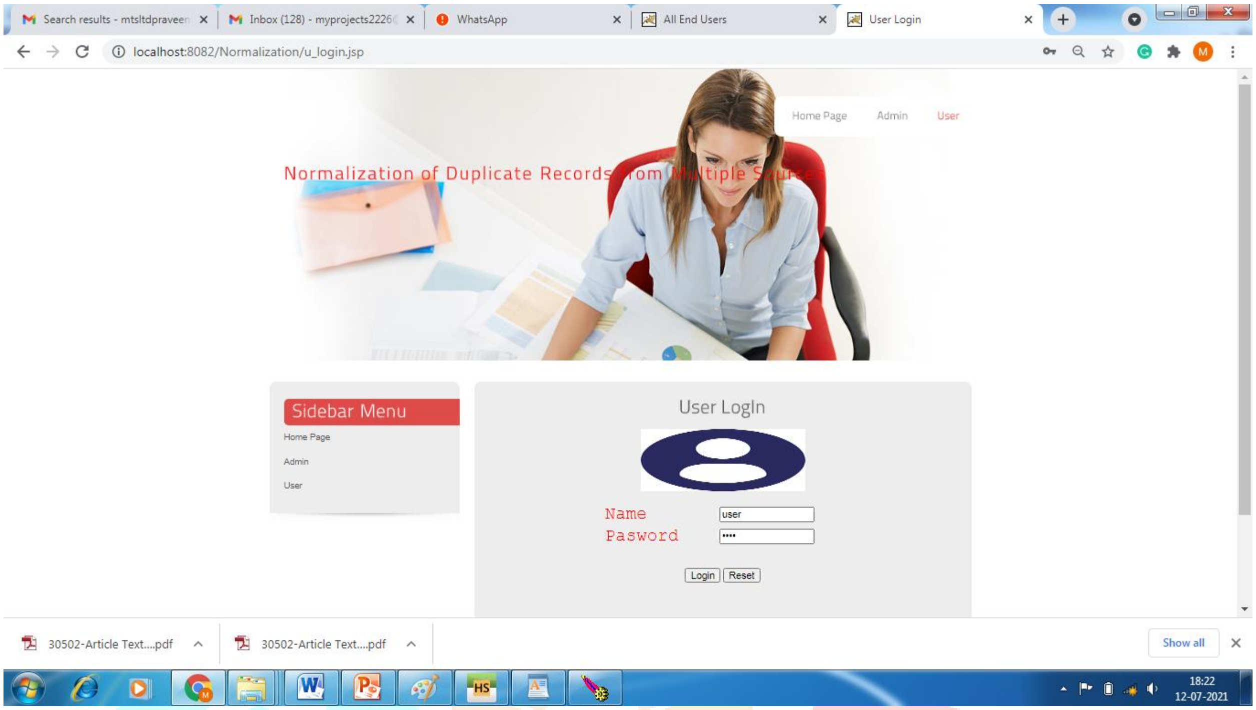
Mobile Number (required)

User Registered Sucessfully

Back Home

Figure . User Try to Register

User Login



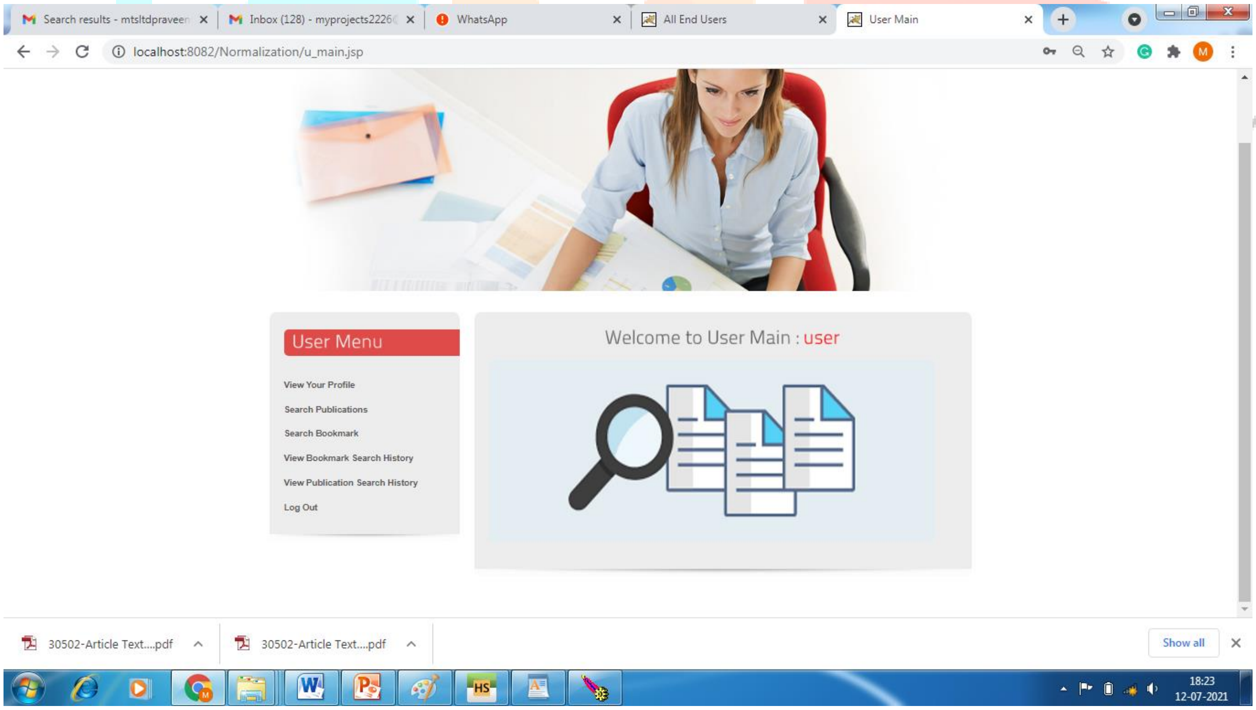
localhost:8082/Normalization/u_login.jsp

Normalization of Duplicate Records from Multiple Sources

User LogIn

Name user

Pasword ••••



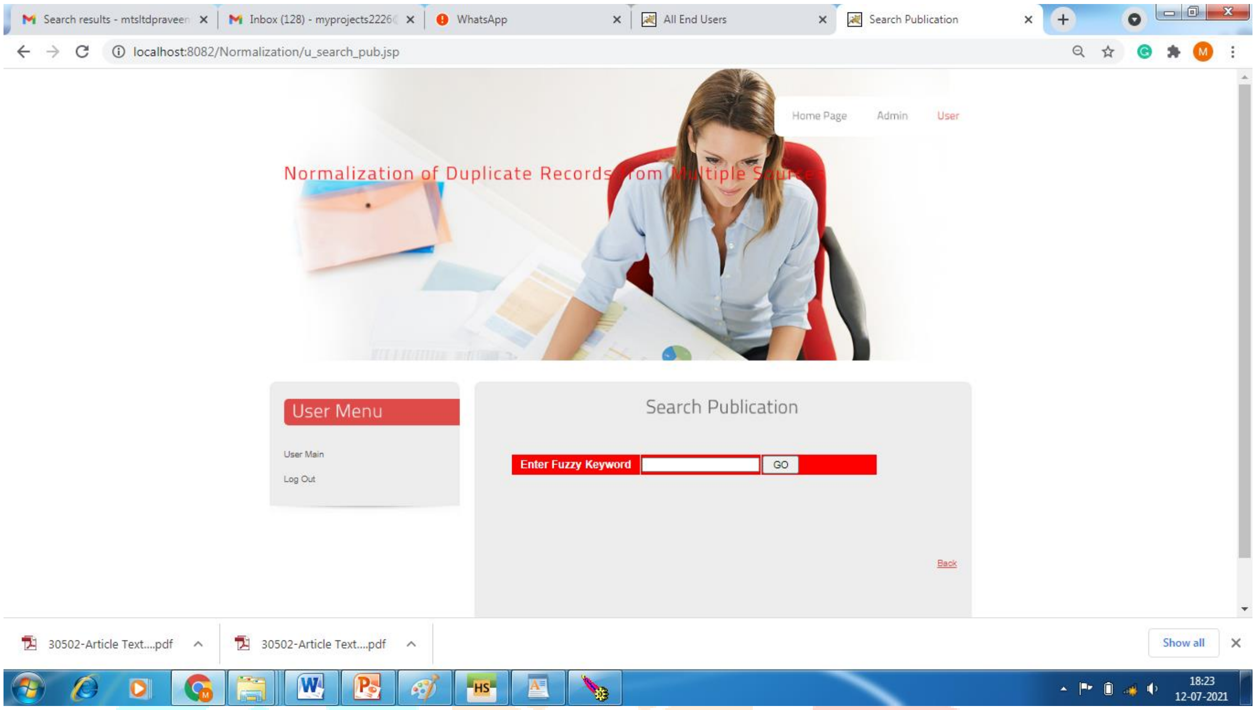
localhost:8082/Normalization/u_main.jsp

Welcome to User Main : user

Search results - mtsltdpraveen
Inbox (128) - myprojects2226
WhatsApp
All End Users
User Main
User Menu
View Your Profile
Search Publications
Search Bookmark
View Bookmark Search History
View Publication Search History
Log Out
30502-Article Text....pdf
30502-Article Text....pdf
Show all
18:23
12-07-2021

Figure . User Try to Login and Enter Main Page

User Search Publication



localhost:8082/Normalization/u_search_pub.jsp

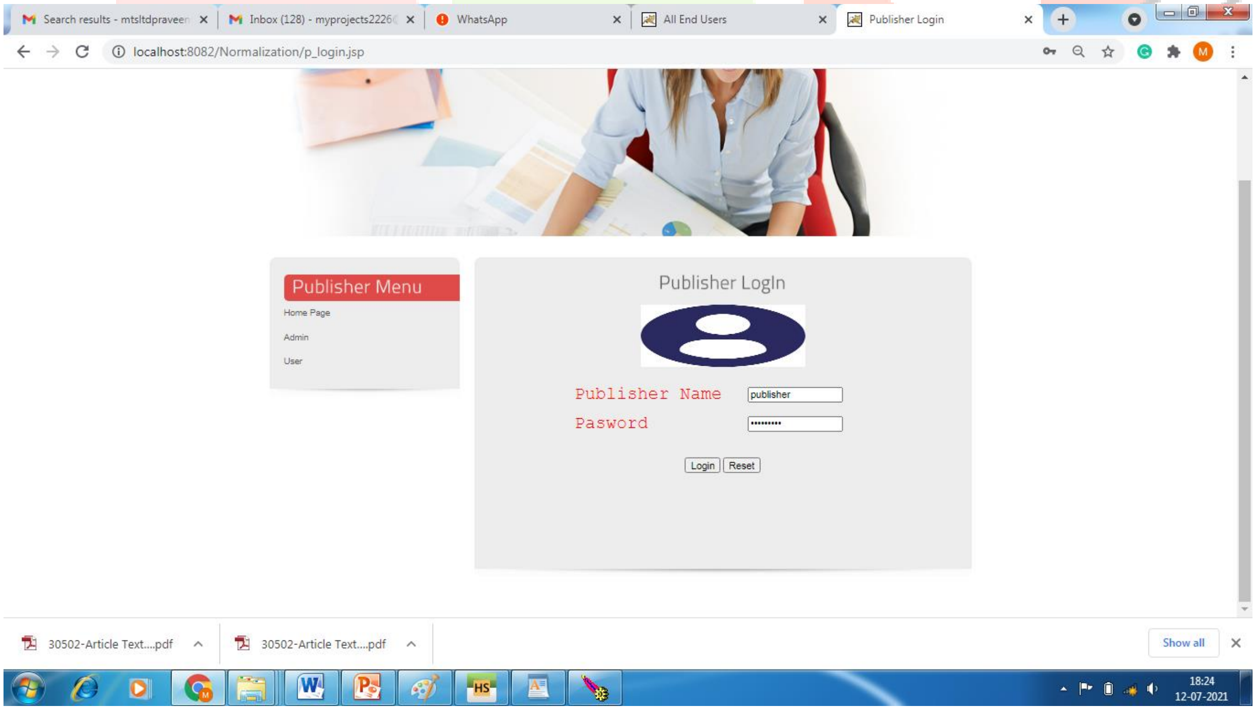
Normalization of Duplicate Records from Multiple Source

Search Publication

Enter Fuzzy Keyword

Figure . User Try to Search Publication

Publisher Login



localhost:8082/Normalization/p_login.jsp

Publisher LogIn

Publisher Name publisher

Pasword •••••••••

Search results - mtsltdpraveen
Inbox (128) - myprojects2226
WhatsApp
All End Users
Publisher Login
Publisher Menu
Home Page
Admin
User
Login
Reset
30502-Article Text....pdf
30502-Article Text....pdf
Show all
18:24
12-07-2021

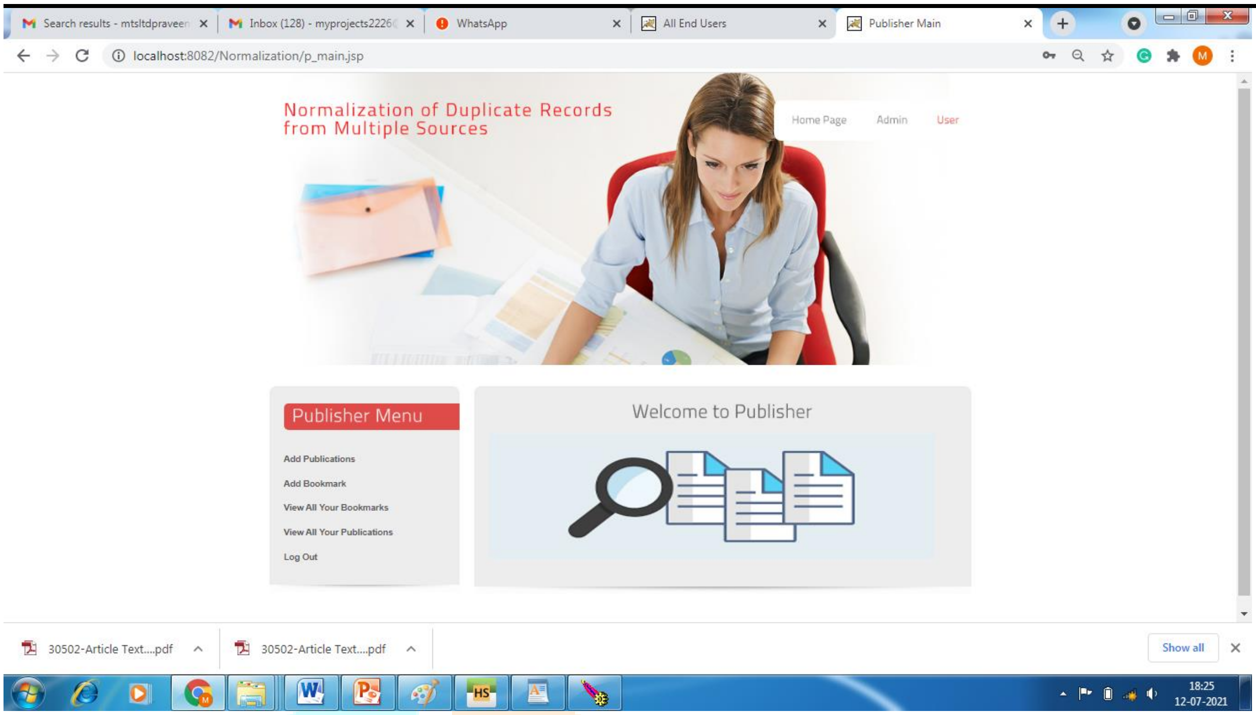

Search results - mtsltdpraveen
Inbox (128) - myprojects2226
WhatsApp
All End Users
Publisher Main
localhost:8082/Normalization/p_main.jsp
Normalization of Duplicate Records from Multiple Sources
Home Page
Admin
User
Publisher Menu
Add Publications
Add Bookmark
View All Your Bookmarks
View All Your Publications
Log Out
Welcome to Publisher
30502-Article Text....pdf
30502-Article Text....pdf
Show all
18:25
12-07-2021

Figure . Publisher Try to login

7. CONCLUSION

In this project, we studied the problem of record normalization over a set of matching records that refer to the same real-world entity. We presented three levels of normalization granularities (record-level, field-level and value component level) and two forms of normalization (typical normalization and complete normalization). For each form of normalization, we proposed a computational framework that includes both single-strategy and multi-strategy approaches. We proposed four single-strategy approaches: frequency, length, centroid, and feature-based to select the normalized record or the normalized field value. For multi strategy approach, we used result merging models inspired from meta searching to combine the results from a number of single strategies. We analyzed the record and field level normalization in the typical normalization. In the complete normalization, we focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values. We implemented a prototype and tested it on a real-world dataset. The experimental results demonstrate the feasibility and effectiveness of our approach. Our method outperforms the state-of-the-art by a significant margin.

8. REFERENCES

[1] K. C.-C. Chang and J. Cho, “Accessing the web: From search to integration,” in *SIGMOD*, 2006, pp. 804–805.

[2] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, “Webtables: Exploring the power of tables on the web,” *PVLDB*, vol. 1, no. 1, pp. 538–549, 2008.

[3] W. Meng and C. Yu, *Advanced Metasearch Engine Technology*. Morgan & Claypool Publishers, 2010.

[4] A. Gruenheid, X. L. Dong, and D. Srivastava, “Incremental record linkage,” *PVLDB*, vol. 7, no. 9, pp. 697–708, May 2014.

[5] E. K. Rezig, E. C. Dragut, M. Ouzzani, and A. K. Elmagarmid, “Query-time record linkage and fusion over web databases,” in *ICDE*, 2015, pp. 42–53.

[6] W. Su, J. Wang, and F. Lochovsky, “Record matching over query results from multiple web databases,” *TKDE*, vol. 22, no. 4, 2010.

[7] H. K¨opcke and E. Rahm, “Frameworks for entity matching: A comparison,” *DKE*, vol. 69, no. 2, pp. 197–210, 2010.

[8] X. Yin, J. Han, and S. Y. Philip, “Truth discovery with multiple conflicting information providers on the web,” *ICDE*, 2008.

[9] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” *TKDE*, vol. 19, no. 1, pp. 1–16, 2007.

[10] P. Christen, “A survey of indexing techniques for scalable record linkage and deduplication,” *TKDE*, vol. 24, no. 9, 2012.

[11] S. Tejada, C. A. Knoblock, and S. Minton, “Learning object identification rules for information integration,” *Inf. Sys.*, vol. 26, no. 8,pp. 607–633, 2001.

[12] L. Shu, A. Chen, M. Xiong, and W. Meng, “Efficient spectral neighborhood blocking for entity resolution,” in *ICDE*, 2011.

[13] Y. Jiang, C. Lin, W. Meng, C. Yu, A. M. Cohen, and N. R. Smalheiser, “Rule-based deduplication of article records from bibliographic databases,” *Database,* vol. 2014, 2014.

[14] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, “Truth finding on the deep web: Is the problem solved?” in *PVLDB*, vol. 6,no. 2, 2012, pp. 97–108.