NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES

V.Sarala [1], D.Vasanthi,

[1]**Assistant professor , MCA DEPT,** Dantuluri Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:- vasanthi.dandu2001@gmail.com
[2]**PG Student of MCA, Dantuluri** Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:- vasanthi.dandu2001@gmail.com

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 hinges upon addressing several big data integration challenges, such as record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there is limited work on creating a uniform, standard record from a group of records corresponding to the same real-world entity. 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 paper, we formalize the record normalization problem, present in-depth analysis of normalization granularity levels (e.g., record, field, and value-component) and of normalization forms (e.g., typical versus complete). We propose a comprehensive framework for computing the normalized record. The proposed framework includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record. 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 and Web tables. 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.Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity, 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.

Literature Survey

The normalization of duplicate records from multiple sources is a critical challenge in data management, particularly in data integration and data cleaning processes. This literature survey aims to summarize significant research contributions in this field, highlighting various methodologies and techniques used to address the problem of duplicate record normalization.

3 IMPLEMENTATION STUDY

EXISTING SYSTEM:

The problem of normalization of database records was first described by Culotta et al.They provided the first attempt to formalize the record normalization problem and proposed three solutions. The first solution uses string edit distance to determine the most central record. The second solution optimizes the edit distance parameters, and the third one describes a feature-based solution to improve performance by means of a knowledge base. Their approach is an instance of typical field value normalization. They did not consider value-component-level normalization. In addition, their gold standard dataset has many instances of unreasonable normalized records.

Proposed System & alogirtham

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.

4.1 Advantages:

The system is very fast due to identification of three levels of normalization granularity such as record, field, and value component.An Exact Duplicate records detection due to Mining Template Collocation-Sub Collocation Pairs

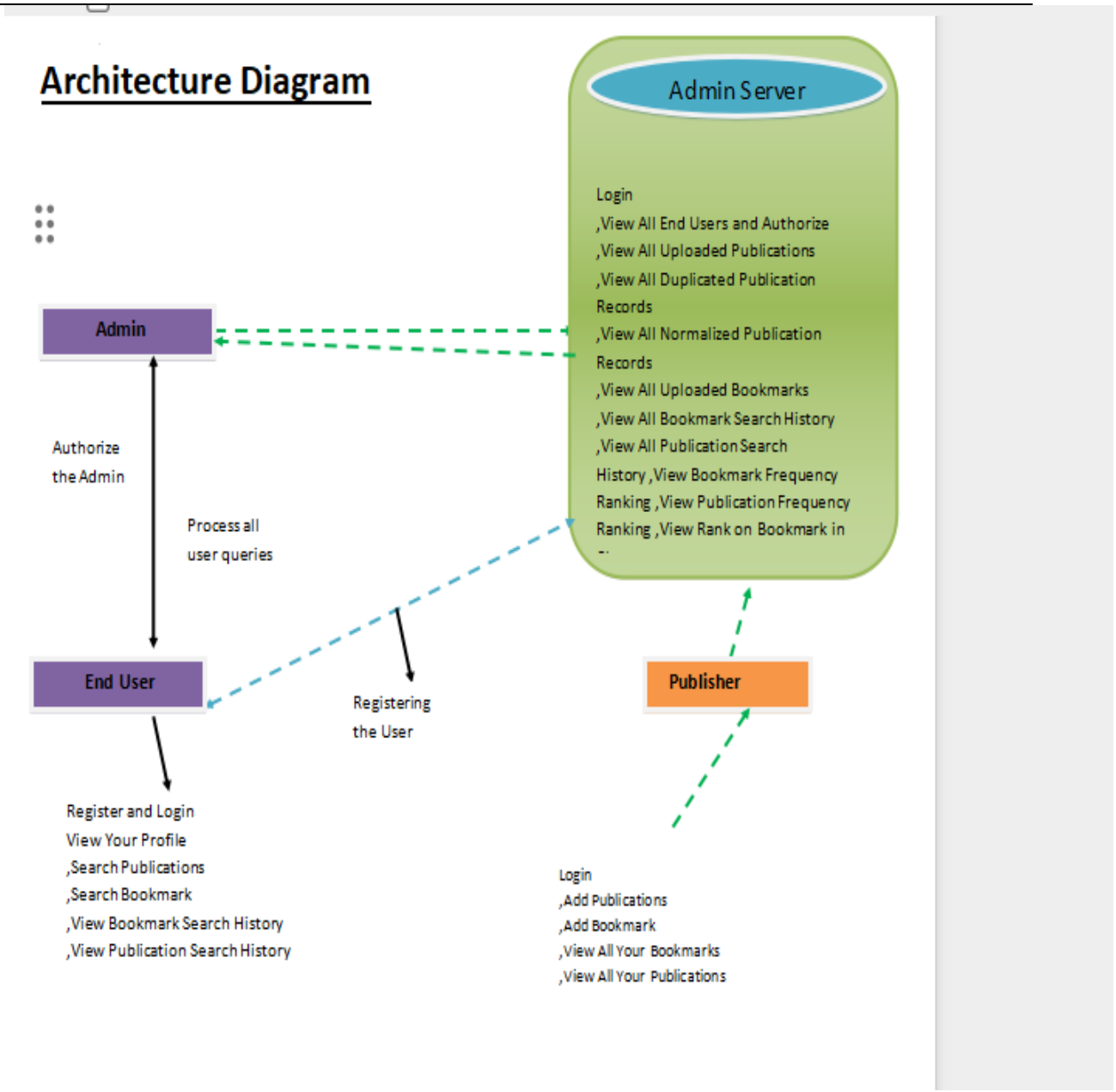

Fig:3.1 System Architecture

IMPLEMENTATION

MODULES:

Admin:

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 View All End Users and Authorize, View All Uploaded Publications, View All Duplicated Publication Records, View All Normalized Publication Records

View All Uploaded Bookmarks, View All Bookmark Search History, View All Publication Search History, View Bookmark Frequency Ranking, View Publication Frequency Ranking, View Rank on Bookmark in Chart, View Rank on Publication in Chart.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

5 RESULTS AND DISCUSSION

HOME SCREEN

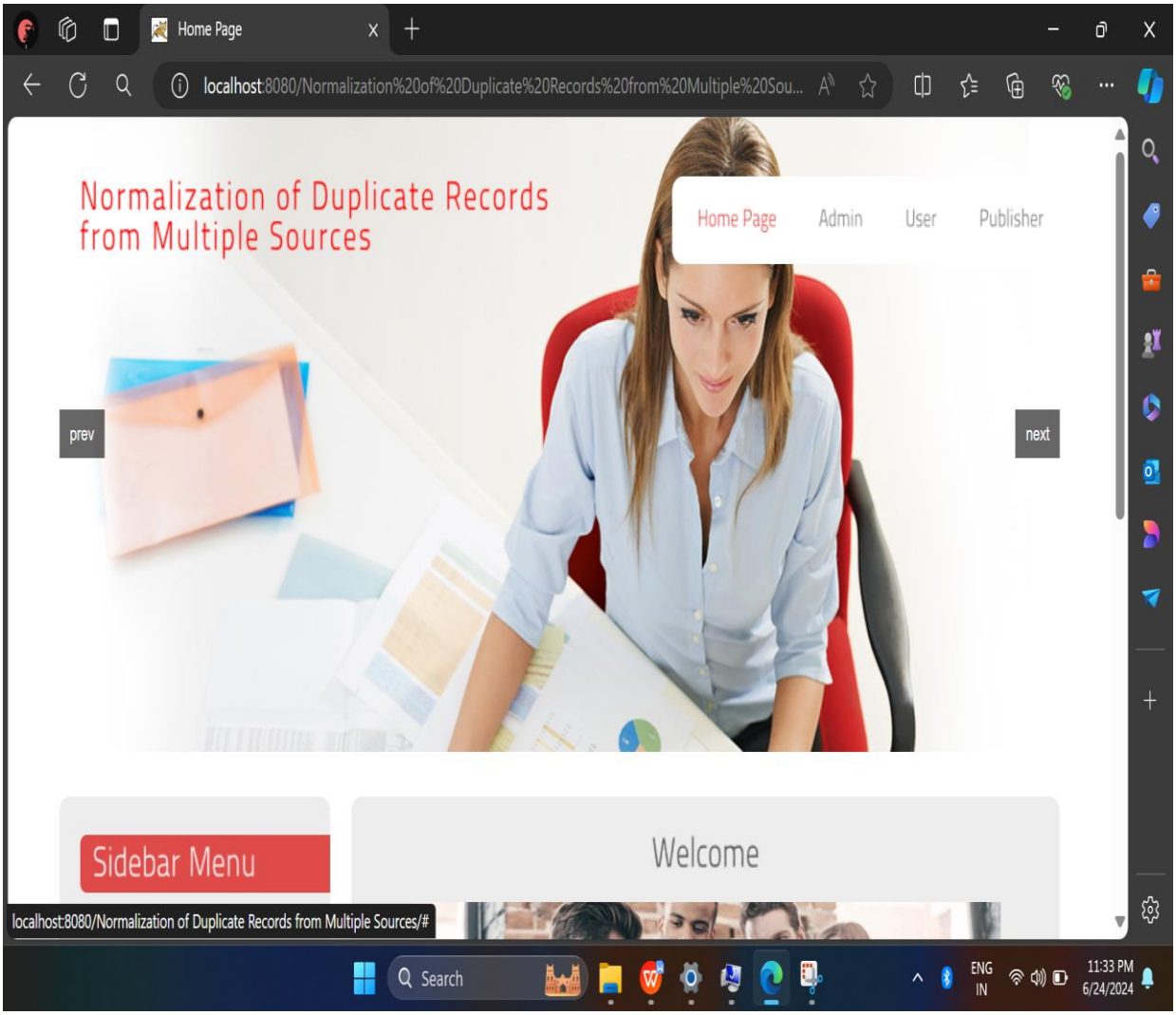

FIG:5.1 HOME SCREEN

59

5.3.2 ADMIN LOGIN SCREEN:

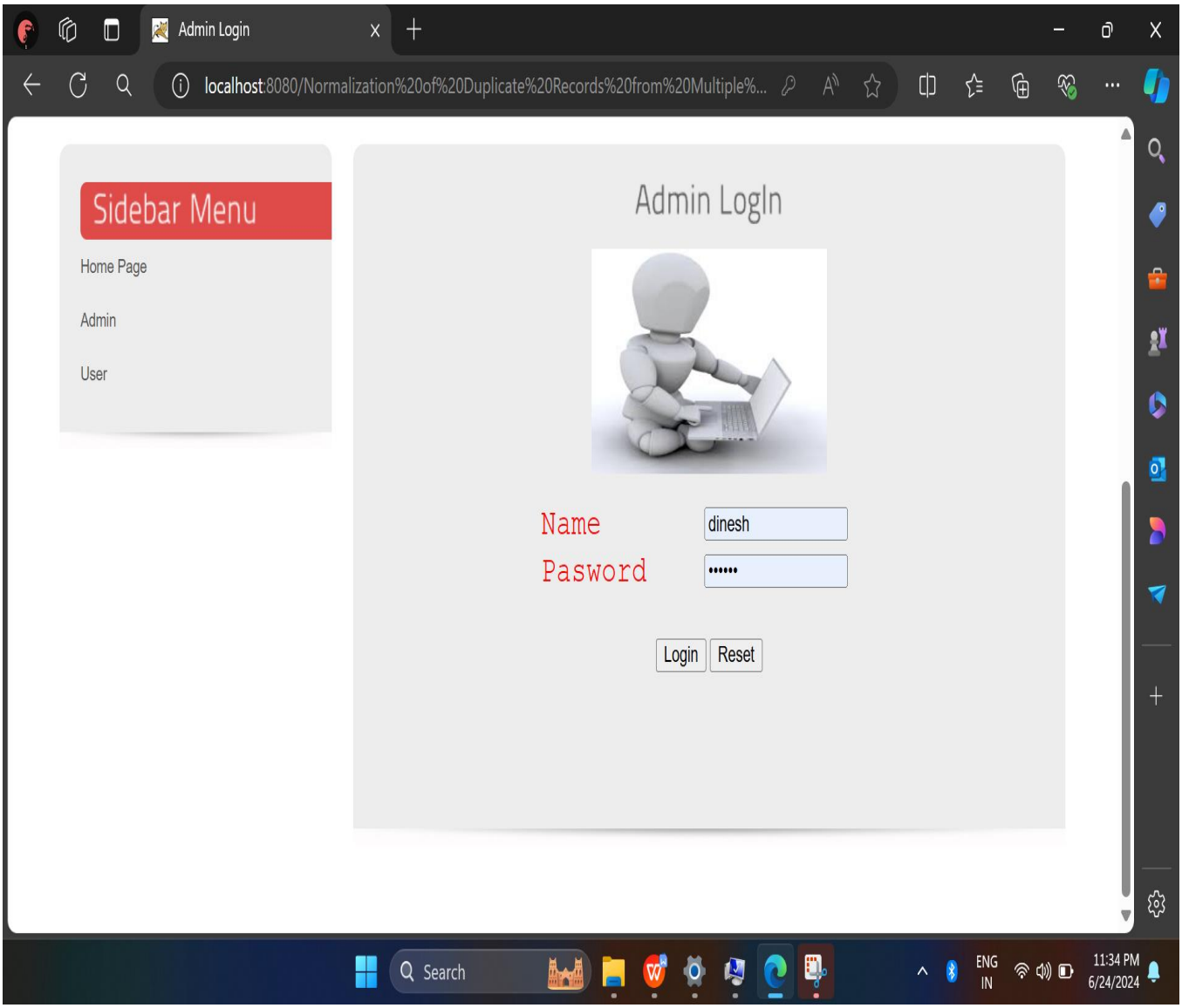

FIG:5.2 ADMIN LOGIN SCREEN

60

5.3.3 ADMIN MENU SCREEN:

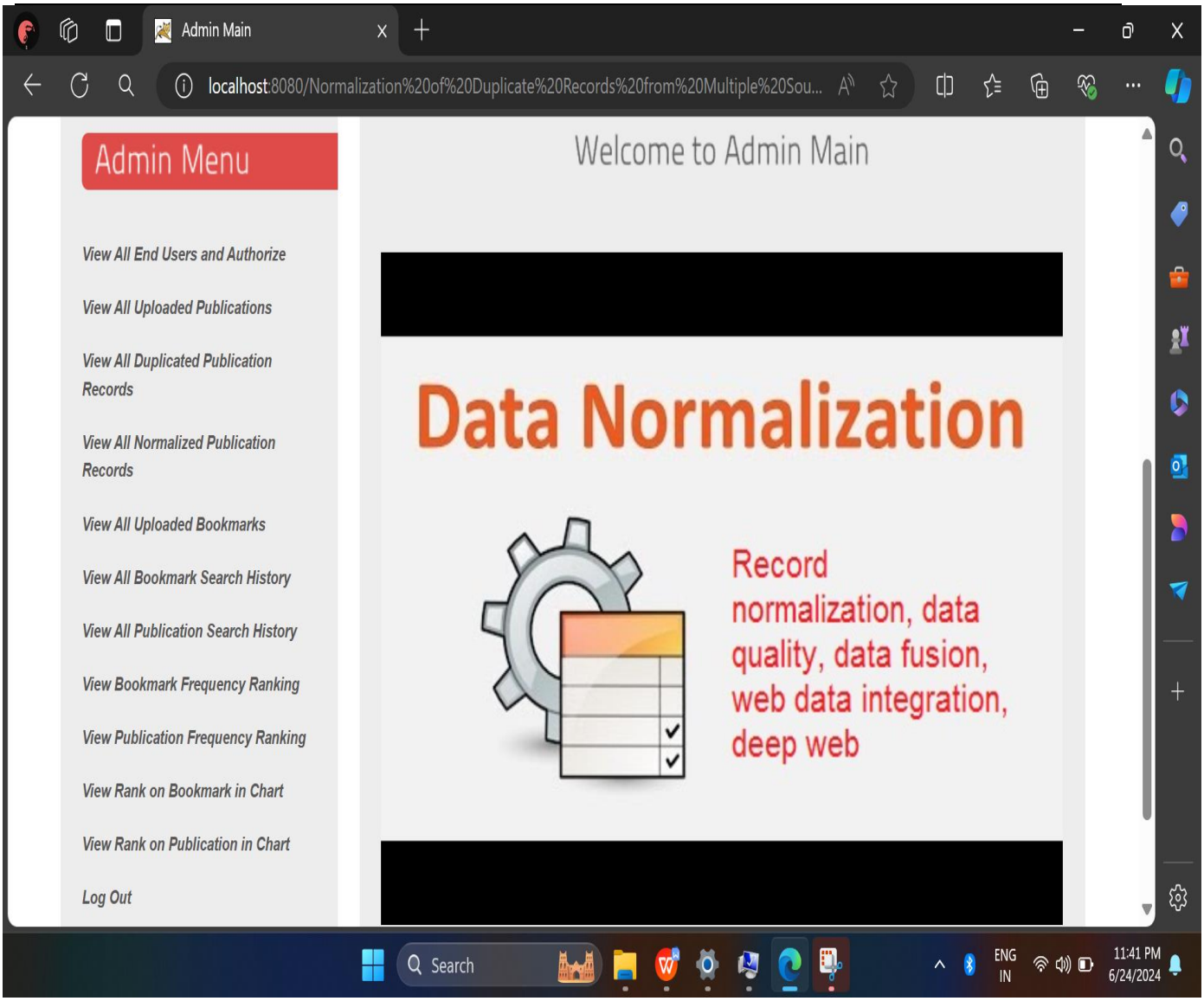

FIG:5.3 ADMIN MENU SCREEN

61

5.3.4 VIEW END USERS SCRREN:

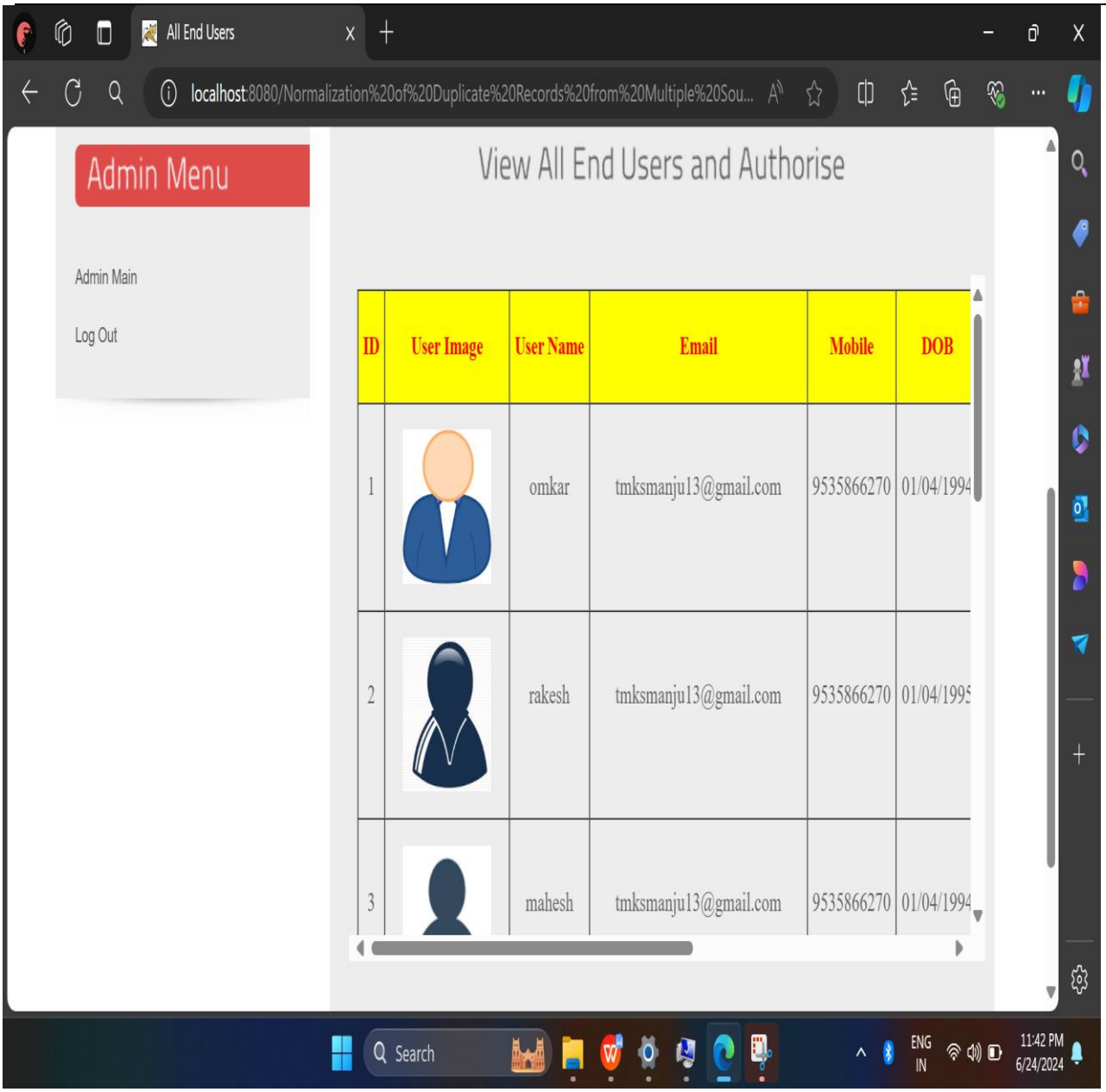

FIG:5.4 VIEW END USER SCREEN

62

5.3.5 VIEW PUBLICATIONS SCREEN:

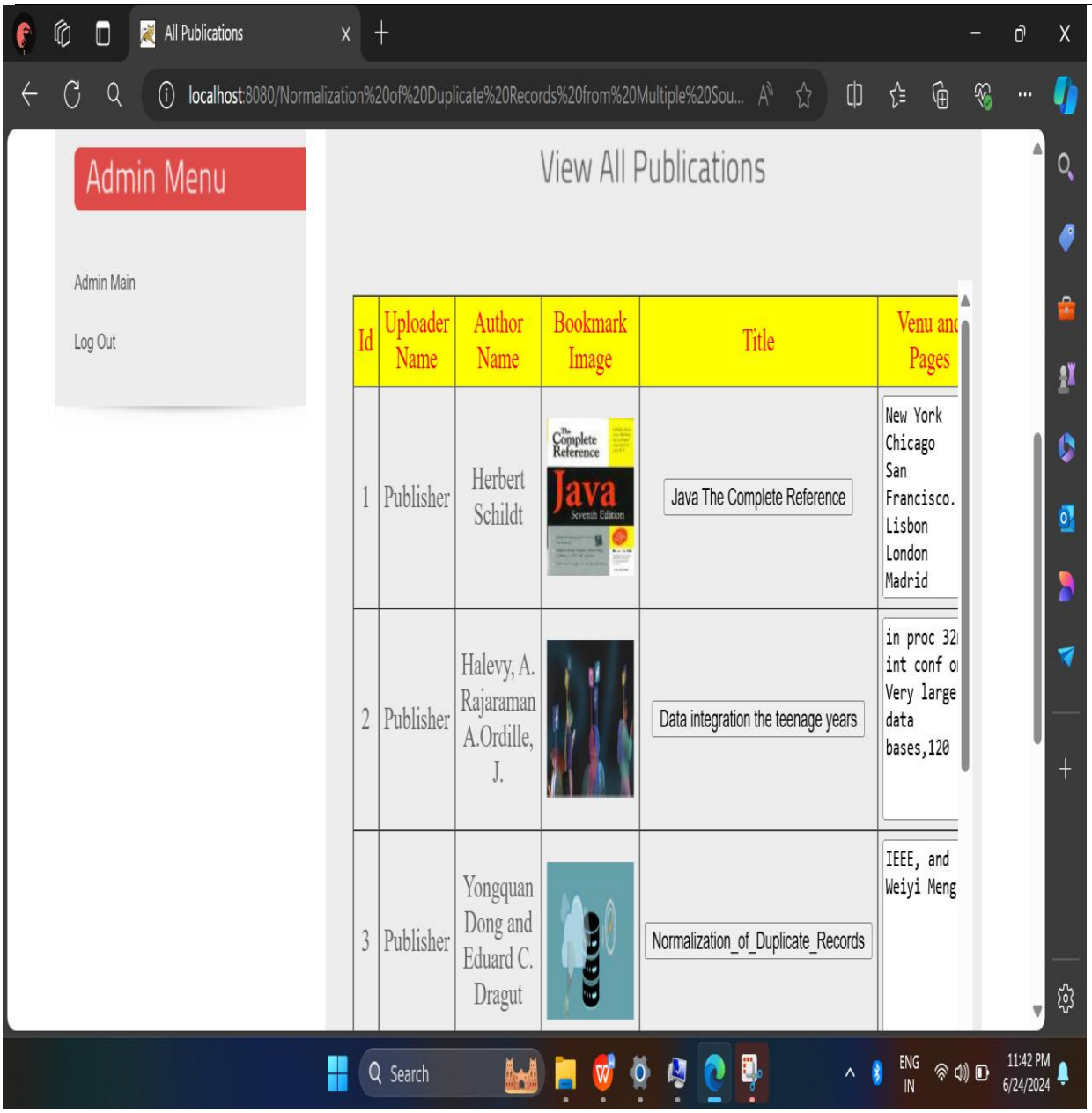

FIG:5.5 VIEW PUBLICATIONS SCREEN

63

5.3.6 VIEW DUPLICATED RECORDS:

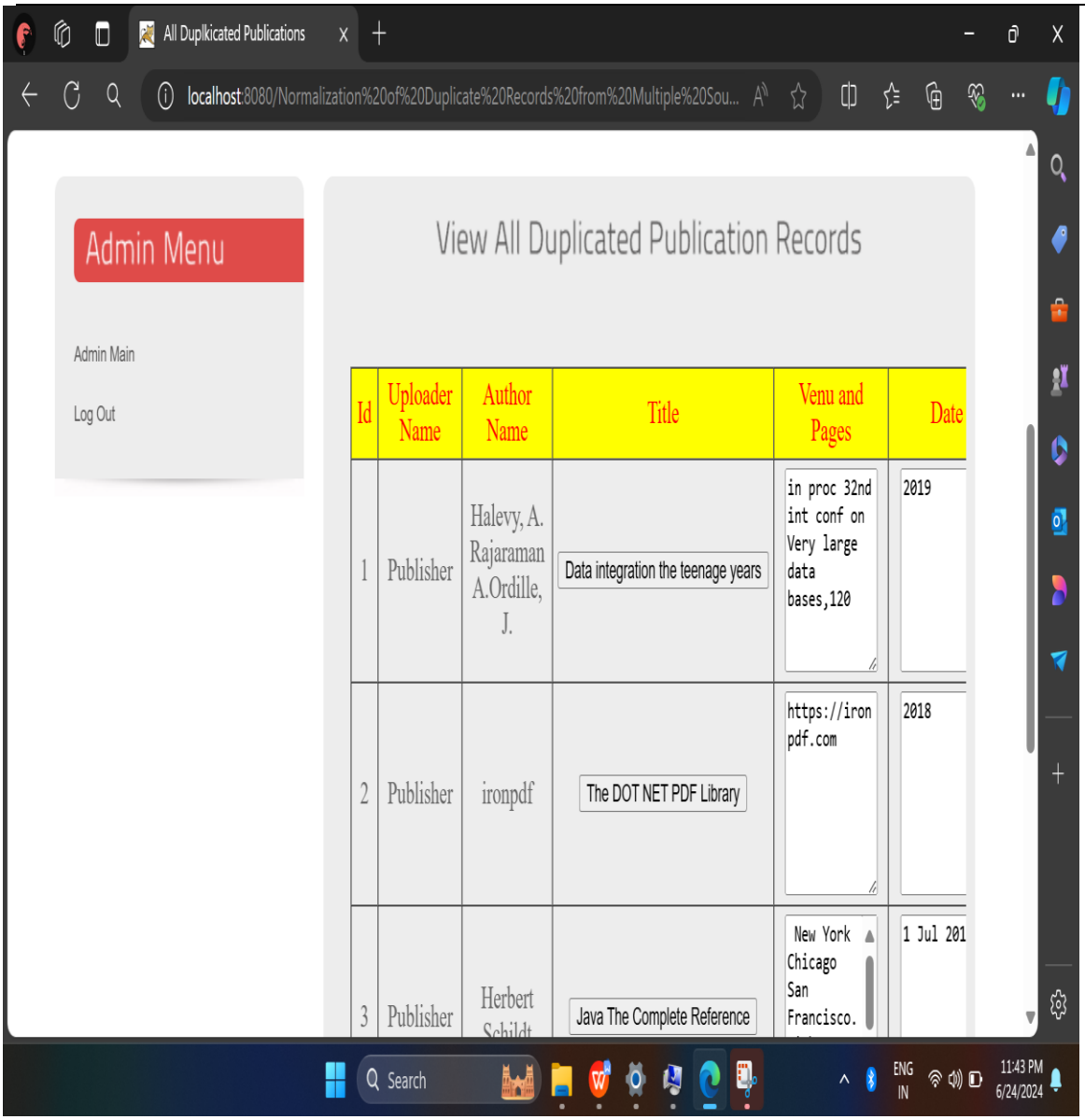

FIG:5.6 VIEW DUPLICATED RECORDS

64

5.3.7 VIEW PUBLICATIONS RANK:

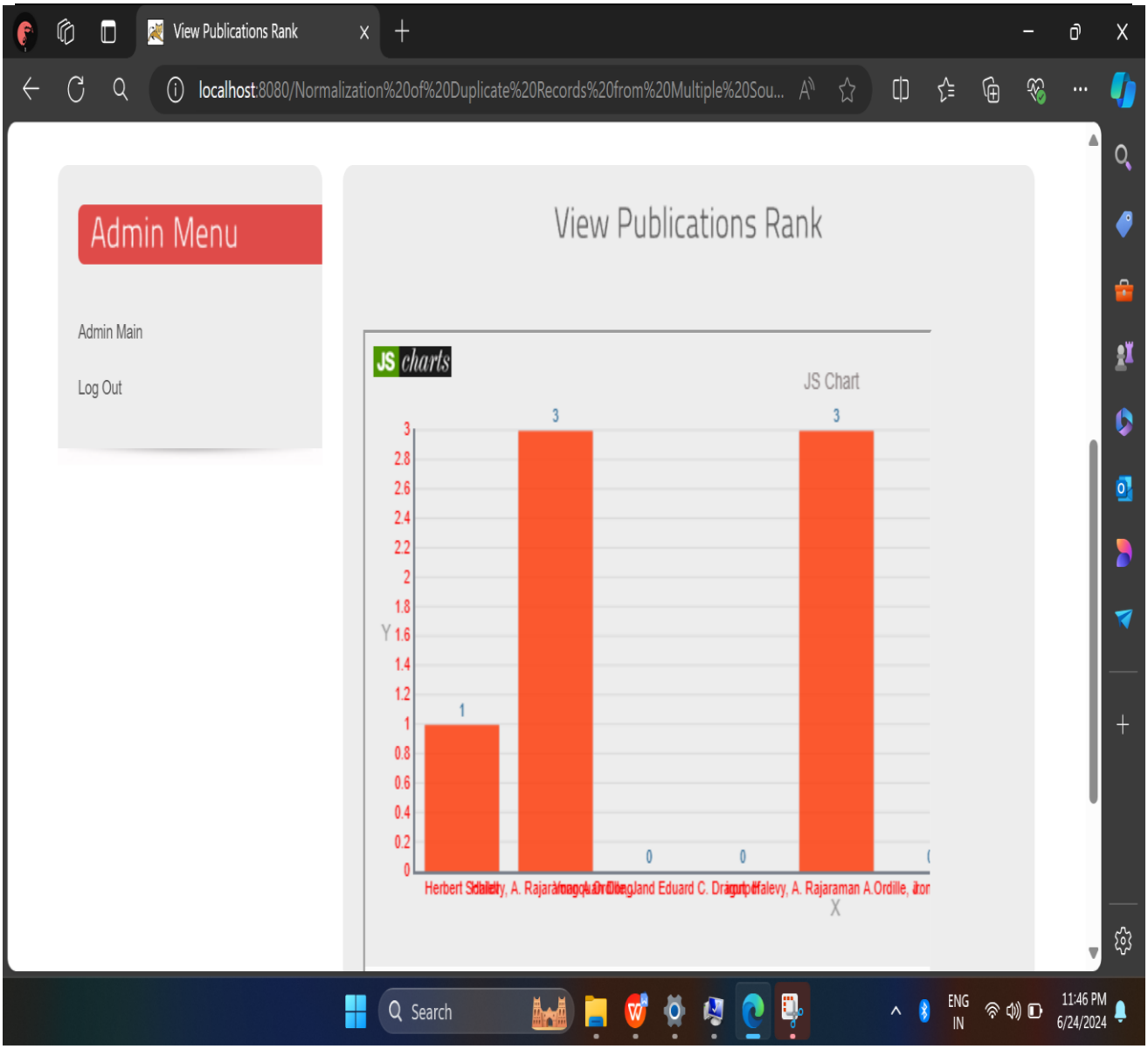

FIG:5.7 VIEW PUBLICATIONS RANK

65

5.3.8 USER LOGIN SCREEN:

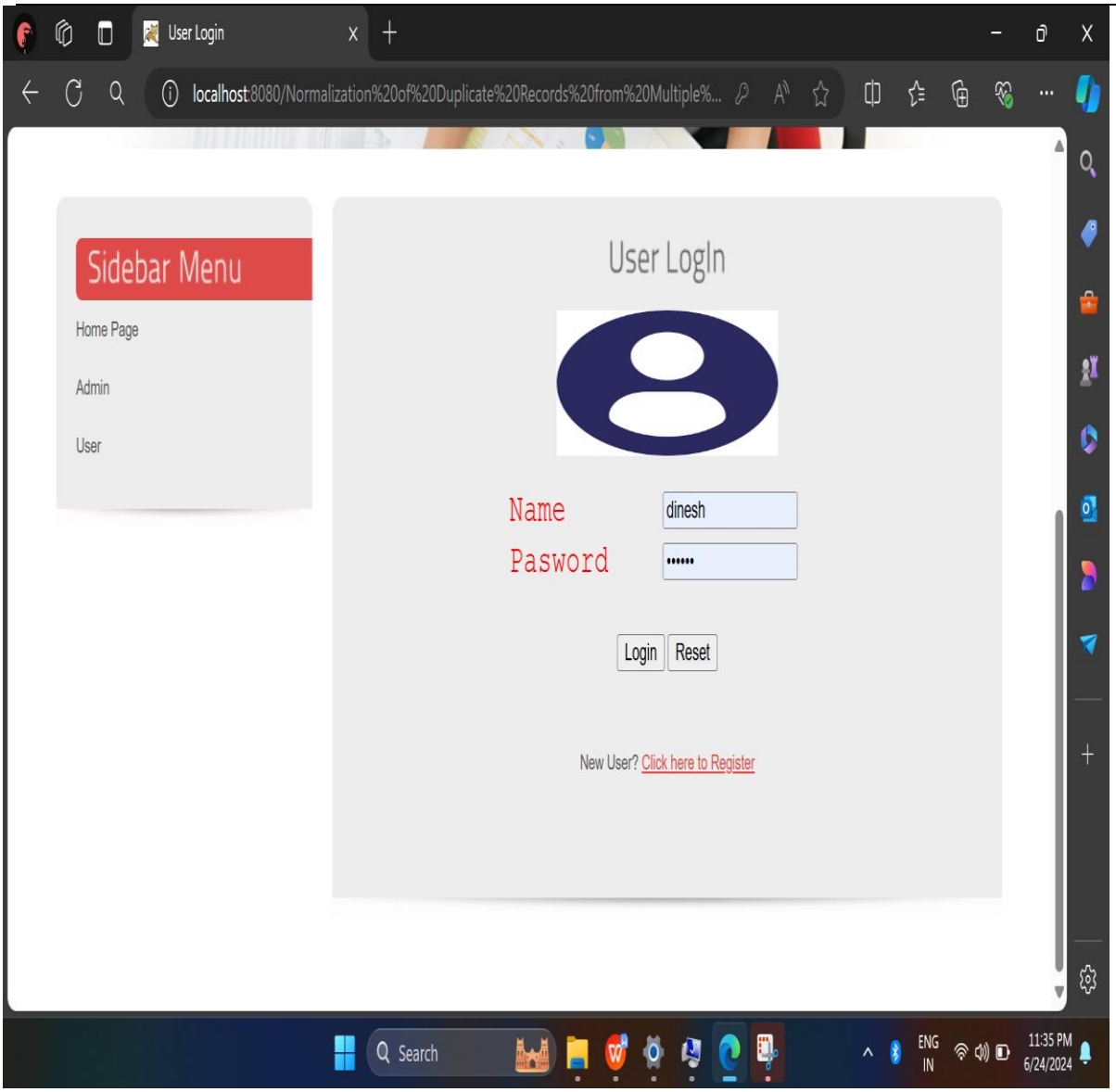

FIG:5.8 USER LOGIN SCREEN

66

5.3.9 USER REISTRATION:

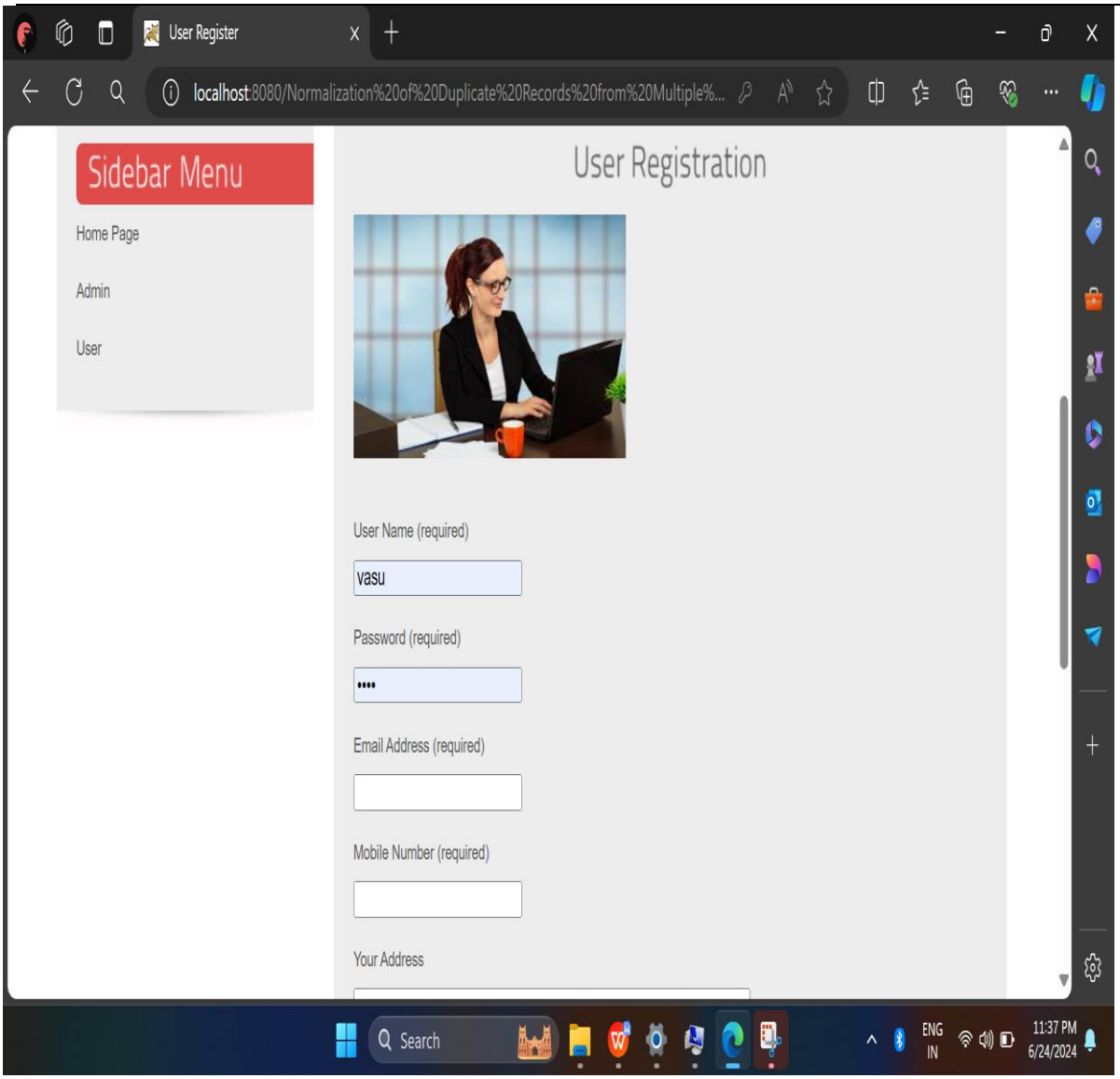

FIG:5.9 USER REGISTRATION

67

5.3.1O ADD PUBLICATION:

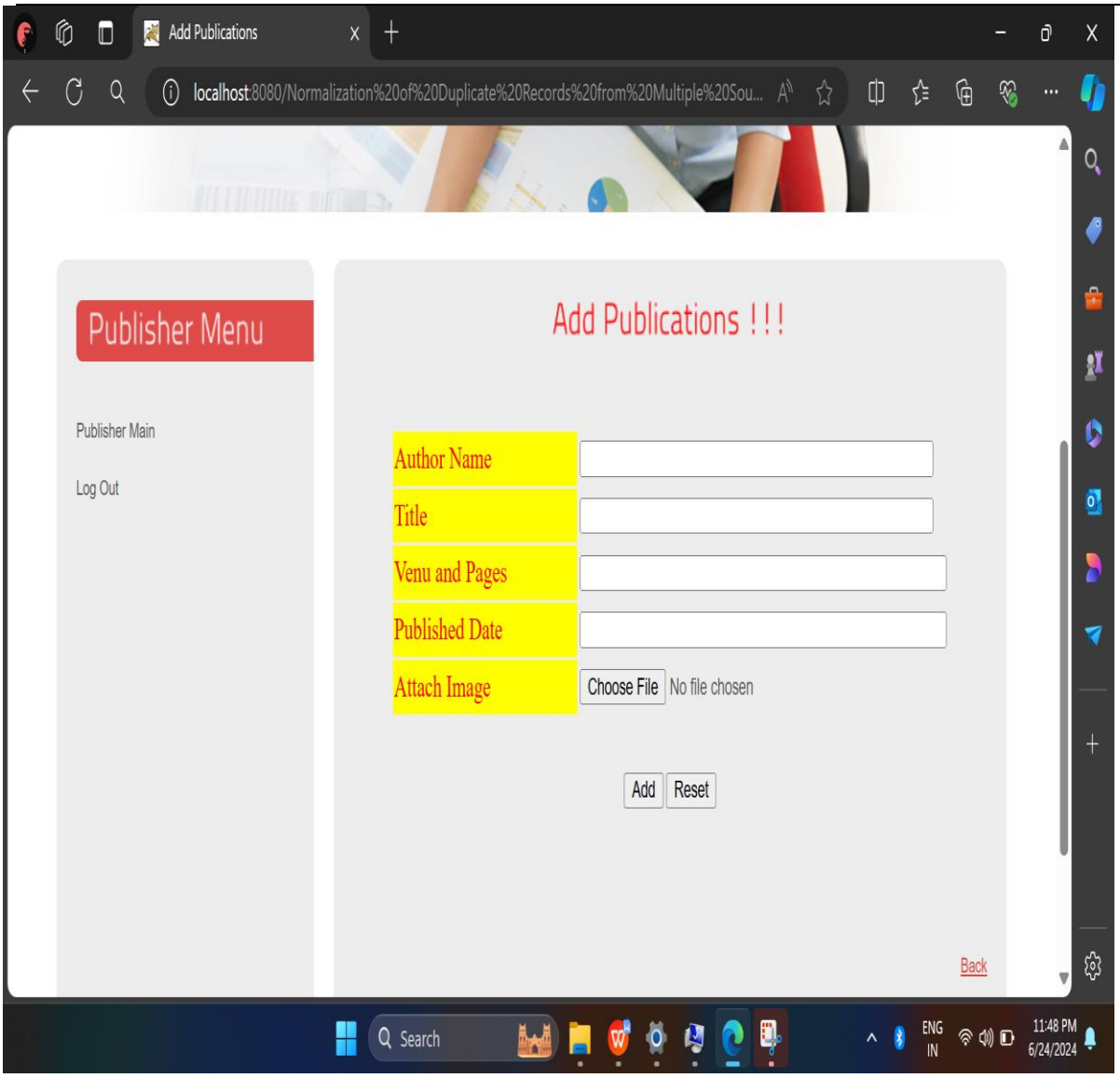

FIG:5.10 ADD PUBLICATION

68

5.3.11 SEARCH PUBLICATION:

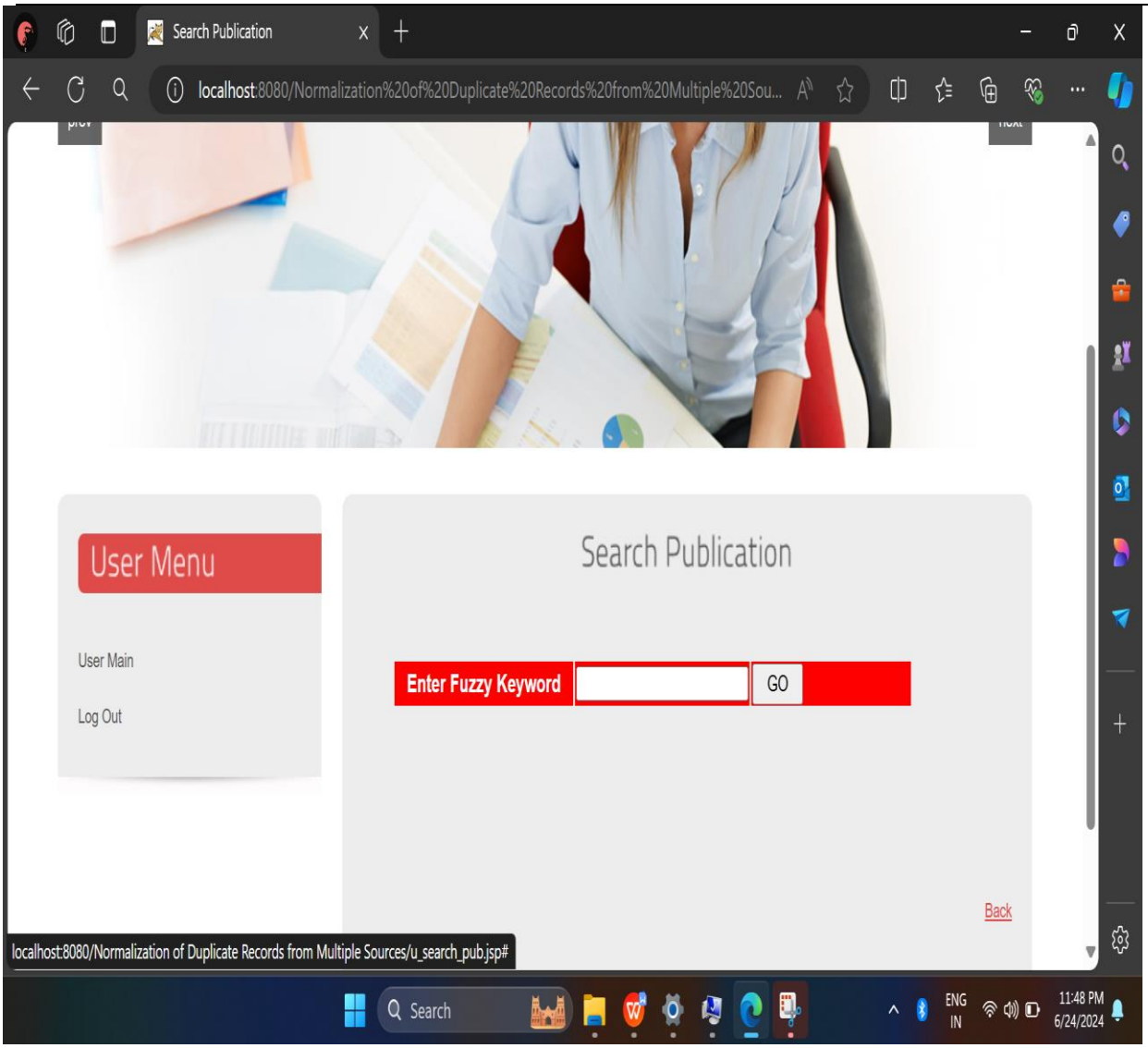

FIG:5.11 SEARCH PUBLICATION

69

5.3.12 PUBLICATION SEARCH HISTORY:

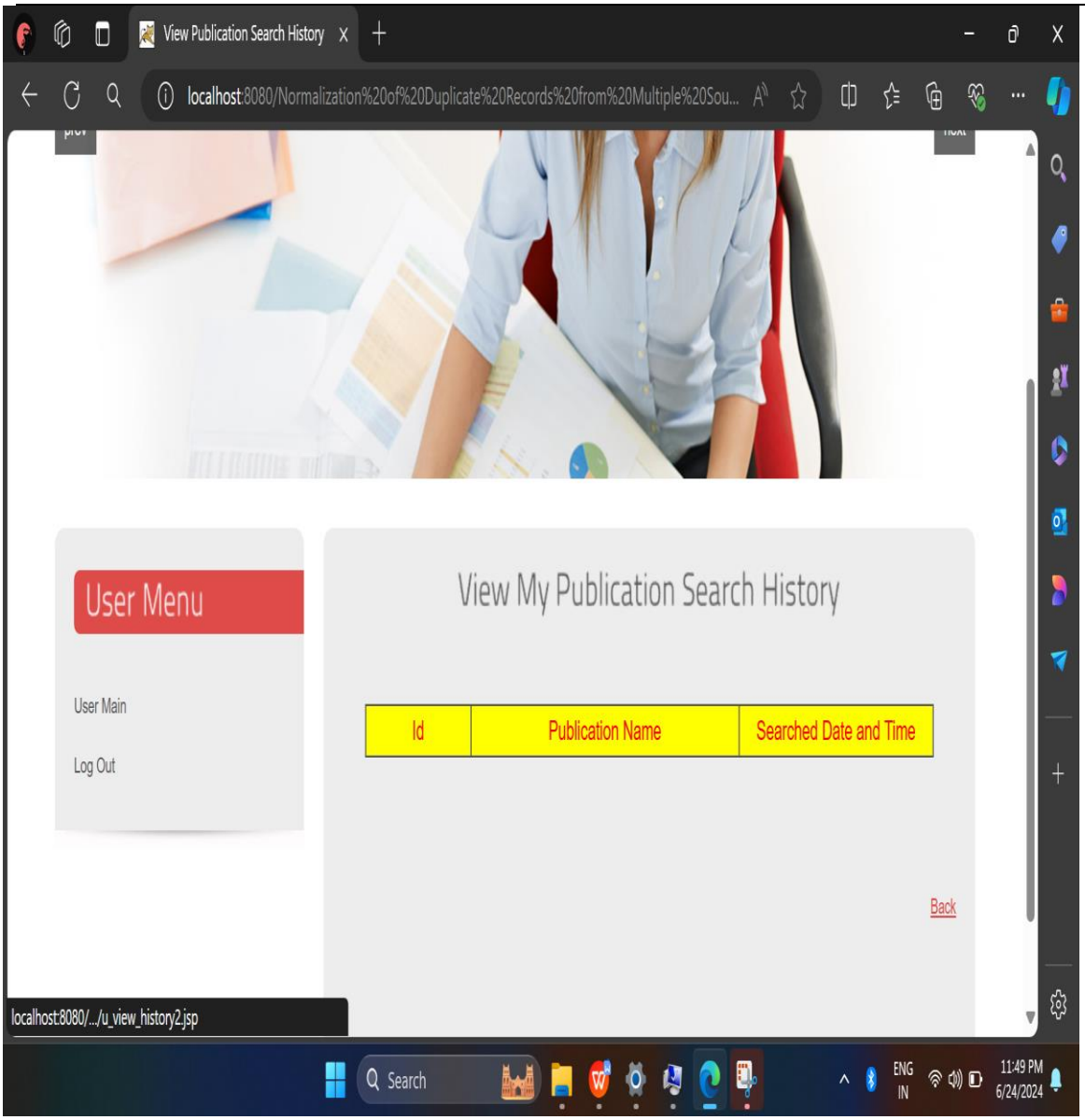

FIG:5.12 PUBLICATION SEARCH HISTORY

6. CONCLUSION AND FUTURE WORK

CONCLUSION

In this paper, 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. In the future, we plan to extend our research as follows.

7. REFRENCES

[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, "Web tables: 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.