Abstract

The purpose of this research was to review, analyze and compare algorithms lying under the empirical technique in order to suggest the most effective algorithm in terms of efficiency and accuracy. The research process was initiated by collecting the relevant research papers with the query of “duplication record detection” from IEEE database. After that, papers were categorized on the basis of different techniques proposed in the literature. In this research, the focus was made on empirical technique. The papers lying under this technique were further analyzed in order to come up with the algorithms. Finally, the comparison was performed in order to come up with the best algorithm i.e. DCS++. The selected algorithm was critically analyzed in order to improve its working. On the basis of limitations of selected algorithm, variation in algorithm was proposed and validated by developed prototype.

After implementation of existing DCS++ and its proposed variation, it was found that the proposed variation in DCS++ producing better results in term of efficiency and accuracy. The algorithms lying under the empirical technique of duplicate records deduction were focused. The
research material was gathered from the single digital library i.e. IEEE. A restaurant dataset
was selected and the results were evaluated on the specified dataset which can be considered
as a limitation of the research. The existing algorithm i.e. DCS++ and proposed variation in
DCS++ were implemented in C#. As a result, it was concluded that proposed algorithm is
performing outstanding than the existing algorithm.

References

   Method for Near-Duplicated Records Detection," in Computational Intelligence and Software
3. M. Rehman and V. Esichaikul, "DUPLICATE RECORD DETECTION FOR DATABASE
   2009 , pp. 333 - 338.
4. X. Mansheng, L. Yoush, and Z. Xiaqi, "A PROPERTY OPTIMIZATION METHOD in
   SUPPORT of APPROXIMATELY DUPLICATED RECORDS DETECTING," in Intelligent
5. Q. Hua, M. Xiang, and F. Sun, "An Optimal Feature Selection Method for Approximately
   Duplicate Records," in Information Management and Engineering (ICIME), 2010 The 2nd IEEE
   International Conference on, Chengdu, 2010.
   in Nirma University International Conference on Engineering, 2013.
7. L. Huang, P. Yuan, and F. Chu, "Duplicate Records Cleansing with Length Filtering and
   Dynamic Weighting," in Semantics, Knowledge and Grid, 2008. SKG '08. Fourth International
9. Z. Wei, W. Feng, and L. Peipei, "Research on Cleaning Inaccurate Data in Production," in
   Service Systems and Service Management (ICSSSM), 2012 9th International Conference on,
   Shanghai, 2012.
10. L. Zhe and Z. Zhi-gang, "An Algorithm of Detection Duplicate Information Based on
11. H., H. Shahriri and Z., A., A. Barforush, "Data Mining for Removing Fuzzy Duplicates
    Using Fuzzy Inference," in Processing NAFIPS '04. IEEE Annual Meeting of the (Volume:1 ),
    2004.
12. W. Su, J. Wang, and H., F. Lochovsky, "Record Matching over Query Results from
    Multiple Web Databases," in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA
    ENGINEERING, 2010.
    Analytic Feature Weighting," in Third International Conference on Advances in Computing and
    Communications, 2013.
14. L., Wan Zhao and Wah, C. N., "Scale-Rotation Invariant Pattern Entropy for
    Keypoint-Based Near-Duplicate Detection," in IEEE TRANSACTIONS ON IMAGE
    PROCESSING, 2009.


Index Terms

Computer Science Information Sciences

Keywords

Duplication Records Detection Algorithm, DCS++, Windowing, Blocking