Journal of Engineering Research

Volume 1 | Issue 1 Article 14

2015

WEB-BASED DUPLICATE RECORDS DETECTION WITH ARABIC LANGUAGE ENHANCEMENT

Azza Abd Al-Elah Higazy,
amany_sarhan@f-eng.tanta.edu.eg

Amany M. Sarhan,

Tarek E. El-Tobely

The Journal of Engineering Research Volume 1 No.1 Faculty of Engineering-Tanta University

WEB-BASED DUPLICATE RECORDS DETECTION WITH ARABIC LANGUAGE ENHANCEMENT

Eng. Azzaabd Al-Elahhigazy, Ass. Prof. Dr. Amany Mahmoud Sarhan and Ass. Prof. Dr. Tarek E. El-Tobely
Computer and Control Engineering Department,
Faculty of Engineering, Tanta University

Abstract

Sharing data between organizations has growing importance in many data mining projects. Data from various heterogeneous sources often has to be linked and aggregated in order to improve data quality. The importance of data accuracy and quality has increased with the explosion of data size. The first step to ensure the data accuracy is to make sure that each real world object is represented once and only once in a certain dataset which called Duplicate Record Detection (DRD). These data inaccuracy problems exist due to due to several factors including spelling, typographical and pronunciation variation, dialects and special vowel and consonant distinction and other linguistic characteristics especially with non-Latin languages like Arabic. In this paper, an English/Arabic enabled web-based framework is designed and implemented which considers the user interaction to add new rules, enrich the dictionary and evaluate results is an important step to improve system's behavior. The proposed framework allows the processing on both single language dataset and bi-lingual dataset. The proposed framework is implemented and verified empirically in several case studies. The comparison results showed that the proposed system has substantial improvements compared to known tools.

Keywords: *Duplicate Record Detection, Data Cleaning, Indexing; Data Integration, Entity* Matching, Soundex, Dictionary Building, Similarity Metrics

1. Introduction

Recently business intelligence and data mining solutions becomes the core of the business management processes and services such as decision making, reporting and statistics. Hence all these processes depend on data insertion and retrieval. Data quality has taken more attention, where the accuracy of this service implementation is affected by the quality of data [59]. Many researches targeting data quality problem and solutions; identifying duplicated records (i.e. more than one record refer to the same real world object) in databases is an essential step to ensure data quality, where errors due to duplicate records can harm the overall decision making process.

In this paper, a study for duplicate record detection (DRD) and current challenges is introduced. In order to perform DRD in an efficient manner, this thesis proposes a web-based framework with enhanced techniques that helps to overcome some of the current challenges that face the available frameworks and developed techniques [22, 50].

Duplicate Record Detection (DRD) is the process of identifying all records that refer to the same real-world object. The DRD could be an independent process to perform some statistics or business intelligence operations, or it could be used to establish the Record Linkage, where records from multiple data sources required to be linked based on record identifier, thus defining all records that refer to the same real-world object is an essential step [23]. When a unique identifier for each entity exists and shared across all database records, DRD process becomes trivial. Data quality problems, such as misspelling during data entry, default values, integrating heterogeneous databases….etc. cause the DRD process to be much more complicated than what appears.

Many researches in Record Linkage/Duplicate Detection have been developed and introduced. Some of them were about providing a complete framework or implementing techniques/algorithms that handle a specific stage in DRD [60].
The general steps for record linkage/duplicate detection [17] are; first is data cleaning and standardization where input data is converted into a well-defined form. Then indexing takes a place to generate pairs of records, as records are grouped based on selected Indexing algorithm to reduce the number of comparisons to be made as much as possible. The final step is the classification of each pair of records into duplicates or non-duplicates based on the similarity value.
Recent research in record linkage has concentrated on improving the Classification step, various classification techniques have been developed such as support vector machine (SVM) [49] and k-means clustering [36]. If record pairs are classified into approximate duplicate, a clerical review process is required where these pairs are manually assessed and classified into duplicates or non-duplicates [17].
Many frameworks are available to perform duplicate record detection in case of having records identified by string data like person name. The available frameworks such as FEBRL and TAILOR, are built based on the general steps for DRD that appeared in [17], which are cleaning and standardization, indexing/blocking, record pair comparison and similarity vector classification. Current challenges of DRD and a comparative study to the most popular indexing techniques will be discussed in later sections in this paper.

3. The Proposed Web-based DRD Framework

There are some limitations in the available frameworks used to perform DRD. Wide variations of typographical representation of some textual information like addresses and persons/places name which appears clearly in non-Latin language such as Arabic, is one from the major problems in DRD. Alternate first names problem appears in western languages, and from performance point of view; the complexity while working with huge amount of data is another issue that must be handled.
In this paper, a web-based Duplicate Record Detection (DRD) framework is designed and implemented to overcome some of the missing features and capabilities in the currently available frameworks. The proposed framework provides black box web service, where there is no need for additional configurations or installations on the client machine. Also it allows the interaction with SME and gives him the ability to add new rules to enhance the system behavior while working in specific type of data, the SME can add the new rules to system temp repository to be trained later by system admin and test the accuracy and decide if these rules can be added to system permanent repository or not.
The framework shown in Figure 1 proposes an enhanced approach to perform DRD over dataset containing single language (English or Arabic) or bi-lingual such as (English and Arabic). The proposed framework was only implemented for English and Arabic languages; however, it can be extended to other languages. Up to our knowledge, this is the first implemented framework that explores the Arabic language and bi-lingual area. It's divided into two main components; first is the web-based frontend which acts as a user interface that allows the direct interaction between the user and the proposed DRD backend, and the second component is the DRD backend. The frontend gives the user the ability to specify the input parameters required to establish DRD process, starting from defining the original dataset and all the parameters used to adjust and customize DRD to match dataset nature. After DRD is performed, the results will be shared with user through the frontend, where he can evaluate and export the results.
The DRD backend receives inputs from the frontend component, and performs the DRD process. The backend is built based on the structure of the general Record linkage/ DRD framework described in [17].

In this work, we introduce modifications in Cleaning and Standardization technique, and Indexing technique keeping the sequence of all DRD flow with no changes. The proposed Cleaning and Standardization component built and optimized using training data, and verified with both benchmark data used in [17], and real data extracted from Egyptian university management information system (MIS).

As a web-based framework, no need for configurations or installations on the client side machine. Another value from building this framework as a web-based is to build an accumulative standardization rules based on the human interaction through the web interface, to improve the systems' behavior through user experience. After reviewing the frequently added rules and testing it using training data, the system expert administration takes a decision if the rules can be added as built-in standardization rules or not.

Through input analyzer component, early prediction of the complexity of DRD can be performed according to the provided data by the user through the frontend, also the representation language/s of data is detected and represented as percentage from all dataset records.After calculating the number of blocks according to BKV, the count of pair of records $\mathbf{Nc}$ will be according to the following equation [46]:$\mathbf{Nc} = \frac{\mathbf{N(N-1)}}{2}$,

Where; N is the number of records inside each block.

Data should be represented in one language before applying standardization rules to grantee the effectiveness of standardization stage.

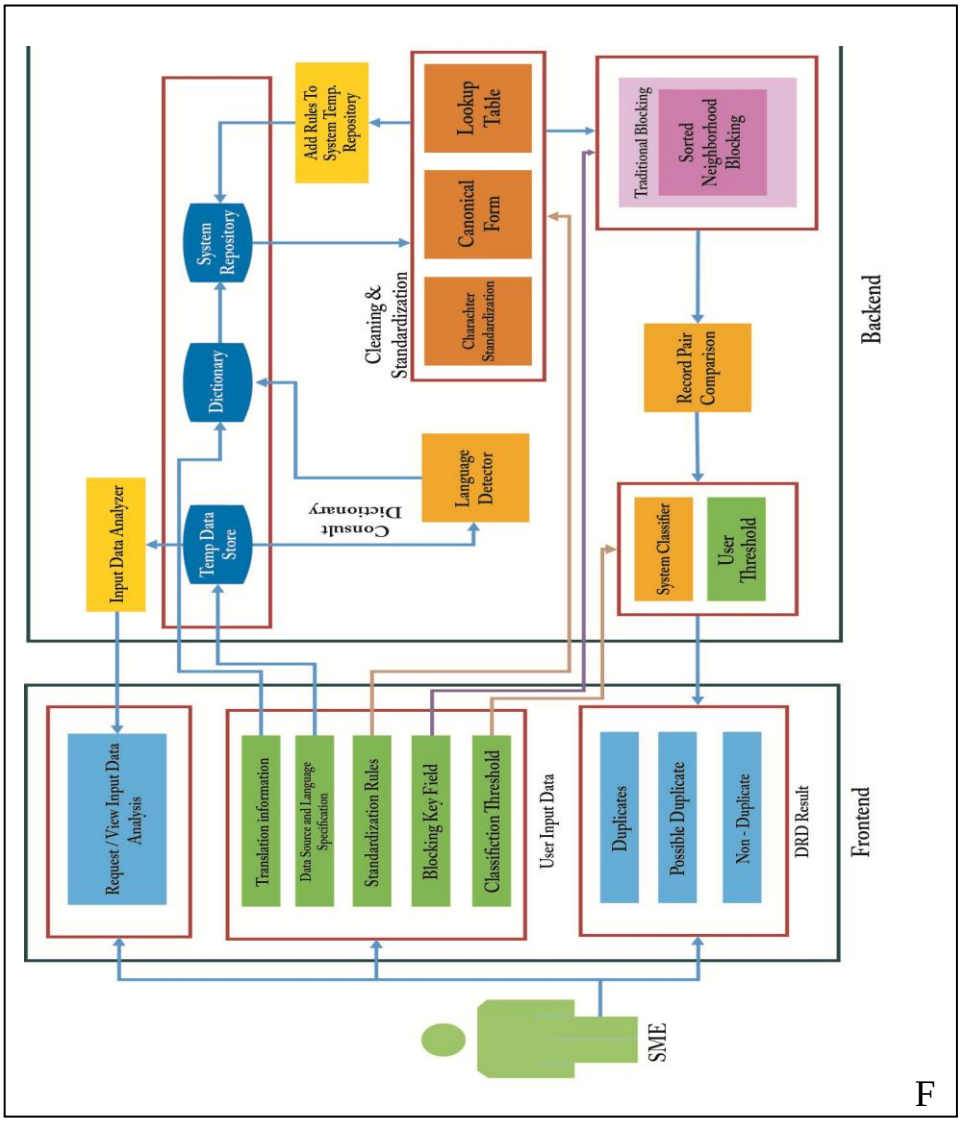

Figure 2 : Proposed Web-Based DRD Framework

TABLE 1. Each pair of records will be classified into duplicates, non-duplicate and possible duplicates, based on the calculated similarity value. The proposed framework uses training data to set the average thresholds value for each type. However, the user can modify this value and set the suitable threshold value according to business requirements.

TABLE 2. The proposed framework is built and verified using labeled training data. The used training data is analyzed to produce rules, which can be used for mapping new data and the same steps are repeated till we get mature results approved by SME. The following approach is followed to extract the training data:

1. We firstly analyze and determine the type of training examples and data to be used.
2. The data is gathered from trusted data sources and it should satisfy the condition of being representative of the real-world use of the function.
3. The learning algorithm is performed on the gathered training data.
4. The accuracy of the learned function is evaluated after parameters' adjustment and learning. The performance of the resulting function is measured on a test set that is separate from the training set.

4. Experimental Results and Discussion

In order to prove the efficiency of the proposed framework, it has been implemented and verified through set of experiments. The proposed web-based DRD is developed using MS SQL server as a backend and ASP.NET to build the frontend, the host environment is Windows Server 2008 R2, 4G RAM and Core i3 processor.

All experiments use real data coming from the following sources:

- Egypt Ministry of Higher Education
- Egyptian university management information system (MIS)
- Supreme Council of Egyptian University (SCU)
- Egyptian Digital Library in SCU
- Egyptian Information Communication and Technology Project (ICTP)

The total Arabic data represented in Arabic is 60,000 records and Arabic data represented in English is 30,000 records.

The data sources are heterogeneous, thus the extracted data has the problem of having multiple representations for the same researcher, without containing unique identifier shared across all data sources. The proposed framework is used to identify all records that refer to the same person (i.e. detect duplicated entities). The next subsections will present the experiments details and the performance analysis.

Experiment 1 (Indexing): Sequential Blocking of English Dataset

In this experiment, the proposed Sequential Blocking technique effectiveness is measured and compared against Traditional Blocking with single BKV, composite BKV and Sorted Neighborhood techniques [17]. We use a dataset that contains 4,000 records for Arabic data represented in English, selected randomly from the scientific research data.

Figure 2 shows a comparison between the number of generated candidate pairs of records by various indexing/blocking techniques. Table 1 shows the number of generated pairs of records in each technique compared to the original number of records in the input dataset. Table 1 shows that

the number of generated record pairs produced by Sorted Neighborhood is 130 times the number of records in the original dataset, where it's decreased to be only 7 times in the proposed Sequential Blocking technique. It is decreased by 17 times compared to Sorted Neighborhood. As shown in Figure 2, the number of candidate pairs of records generated by the proposed technique (Sequential Blocking) is much smaller than its value in the other techniques. This means higher reduction ratio [10].

As shown in Figure 3, this reduction affects the total computational time required for the comparison which will affect the whole DRD process, In Figure 3, the total comparison time in the proposed sequential blocking is downsized to 40% from the Sorted Neighborhood technique.

To ensure that this decrease of computational time did not affect the accuracy of the duplicate record detection process, SME evaluated the DRD results and confirmed the correctness of the obtained results. In clerical review, the results are manually classified by SME to: **True positive (TP):** record pairs classified by machine as duplicates and SME approves the classification results, **True Negatives (TN):** record pairs classified by machine as non-duplicates and SME approves the classification results, **False positive (FP):** record pairs classified by machine as duplicates and SME declines the classification results or **False Negatives (FN):** record pairs classified by machine as non-duplicates and SME declines the classification results.

Blocking Technique	Generated Pairs of Records	Original Dataset Size	No. Candidates to the Original Dataset
Sorted Neighborhood	518,544	4000	130 times
Traditional Single BKV	426,736	4000	107 times
Traditional composite BKV	310,139	4000	78 times
Sequential blocking	29,651	4000	7 times

Table 1: Comparison between various blocking techniques with respect to the number of generated pairs of records compared to the original dataset size.

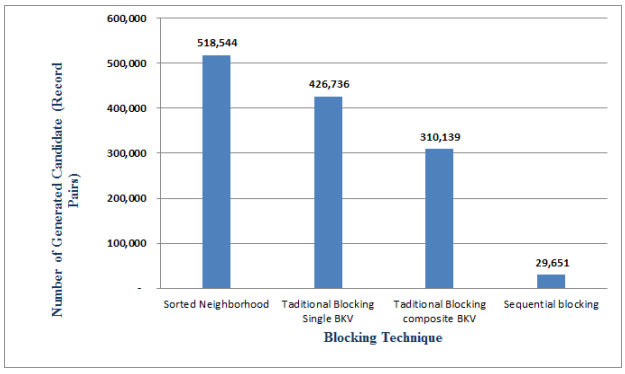

Figure 2: Comparison between number of generated candidate pairs of records by the various indexing/blocking techniques.

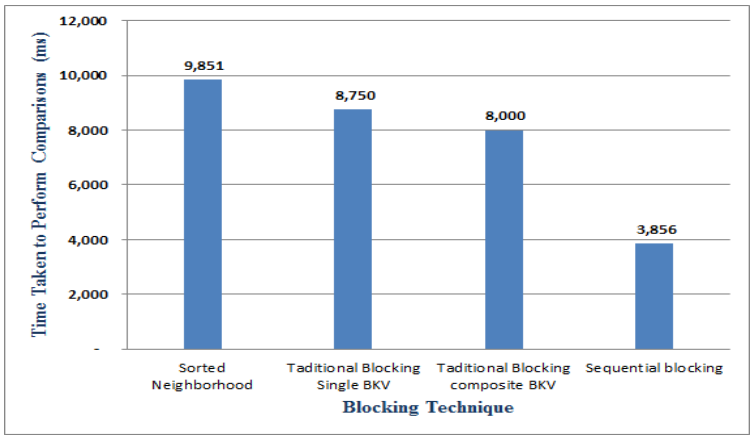

Figure 3: Comparison between comparison computation times required by the various indexing/blocking techniques

In Table 2, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are used to evaluate the results in terms of precision (P) and recall (R). The results in Table 3.11 show that the precision and recall are not affected by the eliminated candidate pairs in the proposed Sequential Blocking technique. This means that all the eliminated candidate pairs were certainly true negatives.

Table 2: Quality and complexity metrics values, for various indexing/blocking techniques

	Traditional Blocking Single BKV	Traditional Blocking Composite BKV	SNH	Proposed sequential Blocking
original dataset	4,000			
No. of generated pairs	426,736	310,139	518,544	29,651
True positives (TP)	166	166	166	166
True Negatives (TN)	426,570	309,973	518,378	29,485
False Positives (FP)	7	4	6	6
False Negatives (FN)	4	2	7	2
P = (TP/(TP+FP))	96%	97%	97%	97%
R = (TP)/(TP+FN))	98%	99%	96%	99%

Experiment 2: Effect of using arabic standardization rules on Arabic dataset

We use a sample of 400 records for scholars' data saved in Arabic language. The dataset contains scholar names, universities and faculties. We use the proposed Sequential Blocking and the classification threshold is tuned to assume that +85% similarity ratios out from Jaro-Winkler distance are candidate duplicated pairs.

First we run the DRD on original dataset without adding the special Arabic standardization (The Arabic Language specific rules). Then we run it again with the proposed standardization approach. The performance analysis for both cases is compared after subject matter expert verification as shown in Table 4.

Table 4: Effect of using Arabic adjustment extensions on the DRD performance analysis for Arabic dataset.

Quality metric	Standard DRD	Proposed Standardization approach
No. of generated pairs	513	513
True positives (TP) pairs	38	54
True Negatives (TN) pairs	77	77
False Positives (FP) pairs	2	2
False Negatives (FN) pairs	19	3
Precision (TP/(TP+FP)) pairs	95%	96%
Recall (TP)/(TP+FN)) pairs	66%	94.7%

It is clear from Table 4 that using the Arabic language adjustment extension caused the true positives to be increased dramatically from 38 to 54 record pairs, the false negatives has been decreased from 19 to 3 record pairs only and the true positive rate (Recall) for the machine has been increased from 66% to 94.7%. This means that the duplicate record detection process quality has been improved substantially.

Experiment 3: Proposed Framework Features Compared to FEBRL

In this experiment, a comparison between the features allowed in the proposed web-based DRD framework and FEBRL is illustrated. This comparison appears in [60] as a part from their study for the available frameworks that perform DRD. The major advantages of the proposed DRD framework over FEBRL are: the availability of enhancing system behavior through real experiments by the users, allowing bi-lingual processing and introducing the sequential blocking technique instead of the current available indexing techniques. However, the proposed framework can be extended to cover more distance functions and indexing techniques as FEBRL does, as the proposed framework was built as proof of concept version.

Table 5 : Comparison between the proposed Web-Based framework and FEBRL

Available Features	Web-based DRD	FEBRL
Unicode support and Language detection algorithm	Yes	No
Similarity Functions	Jaro, Levenshtein	Winkler, Jaro, Q-gram, Positional Q-gram, Skip-gram, Edit distance, Bag distance, Damerau-Levenshtein, Smith-Waterman, and some other functions
Blocking	sequential Blocking, Traditional, SNH, No Blocking	full index, blocking index, sorting index, q-gram index, canopy clustering index, string map index, suffix array index, big match index and deduplication index.
Clerical Reviews Tool	Yes	No
Dictionary building and searching	Yes	No
Metrics Evaluation (TP, Accuracy, Precision, RR)	Yes	No
Lookup	Yes	Yes
Displaying record pair comparison results with details	Enhanced	Partial
Displaying classifier inputs, outputs to trace classifier.	Yes	Yes

5. Conclusion

Duplicate Record Detection (DRD) is an important step in cross-enterprise integration and data mining applications. In this chapter, a web-based framework for Arabic/English DRD is proposed, implemented and verified with enhanced indexing/blocking algorithm to ensure efficiency, where the number of generated candidates is reduced while the ability for true positives prediction (Precision) and false negative prediction (Recall) is saved. To enhance DRD accuracy, a new approach for data cleaning and standardization is provided to perform DRD on dataset contains data represented in Arabic or represented in English, also it provides the ability to unify language of the dataset if it contains English and Arabic at the same time, thus the proposed framework provides a technique to build and enhance dictionary. The proposed framework helps in minimizing the time taken to perform DRD and consequently decrease the effort needed by the user to evaluate the accuracy of the retrieved results.

As long as the framework intends to provide a generic service not limited to specific type/nature of data, it provides the ability to adjust the standardization rules and dictionary during runtime. Also the SME can add the adjustment rules to system temporary data store to be checked later by the system administrator, then the system administrator trains the data using the added rules and decides whether to add it to the system repository or not.

6. References

[1] S.F. Altschula, W. Gisha, W. Millerb, E.W. Meyersc, and D.J. Lipmana, "Basic Local Alignment Search Tool," J. Molecular Biology, vol. 215, no. 3, pp. 403-410, Oct. 1990.

[2] Syed Uzair Aqeel, Steve Beltzel, Eric Jensen, David Grossman and Ophir Frieder, "On the Development of Name Search Techniques for Arabic," American Society for Information Science and Technology, Wiley Periodicals, vol. 57, p. 728-739, 2006.

[3] T. Bachteler, J. Reiher and R. Schnell, "Similarity Filtering with Multibit Trees for Record Linkage," Working Paper WP-GRLC, German Record Linkage Center, Feb. 2013.

[4] R. Baeza-Yates and G.H. Gonnet, "A New Approach to Text Searching," Communications of the ACM, vol. 35, no. 10, pp. 74-82, Oct. 1992.

[5] R. Baxter, P. Christen, and T. Churches, "A comparison of fast blocking methods for record linkage," In Proceeding of ACM SIGKDD'03 workshop on Data Cleaning, Record Linkage and Object Consolidation,Washington DC, pp. 25–27,3-5 Aug., 2003.

[6] M. Bilenko, R.J. Mooney, W.W. Cohen, P. Ravikumar, and S.E. Fienberg, "Adaptive Name Matching in Information Integration," IEEE Intelligent Systems, vol. 18, no. 5, pp. 16-23, Sept. 2003.

[7] V.R. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic Segmentation of Text into Structured Records," In Proceeding of ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA, USA, pp. 175- 186, May 21 - 24, 2001.

[8] A. Chatterjee and A. Segev, "Data Manipulation in Heterogeneous Databases," ACM SIGMOD Record, vol. 20, no. 4, pp. 64-68, Dec. 1991.

[9] P. Christen, T. Churches, M. Hegland, H. Dai, R. Srikant, and C. Zhang, "Febrl: a parallel open source data linkage system," Advances in Knowledge Discovery and Data Mining, Springer, p. 638-647, 2004.

[10] P. Christen and K. Goiser," A Comparison of Personal Name Matching: Techniques and Practical Issues," in proceeding of Data Mining Workshops, ICDM Workshops, 2006.

[11] P. Christen, K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, Studies in Computational Intelligence, Springer, vol. 43, , pp. 127-151, 2007.

[12] P. Christen, "Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification," In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, pp. 151-159, Aug. 24 - 27, 2008.

[13] P. Christen, et al., "Automatic Training Example Selection for Scalable Unsupervised Record Linkage," In Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Springer, Osaka, Japan, pp. 511-518, May 20-23, 2008.

[14] P. Christen, "Febrl - a Freely Available Record Linkage System with a Graphical User Interface," In Proceeding of the 2nd Australasian Workshop on Health Data and Knowledge Management, Wollongong, Australia, vol. 80, pp. 17 -25, Jan. 21-23, 2008.

[15] P. Christen. "Febrl: An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface," In Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08), Las Vegas, USA, pp. 1065–1068, Aug. 24-27, 2008.

[16] P. Christen. "Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System," ACM SIGKDD Explorations Newsletter, vol. 11,

no. 1, p. 39-48, June, 2009.

[17 P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp.1537-1555, September 2012 .

[18] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration," In Proceeding of the ACM SIGKDD'02, Edmonton, pp. 475–480, 2002.

[19] W.W. Cohen, H. Kautz, and D. McAllester, "Hardening Soft Information Sources," In Proceeding of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00), Boston, MA, USA, pp. 255-259, Aug. 20 - 23, 2000.

[20] Dan Wu and Daquing He, "Exploring The Further Integration of Machine Translation in English-Chinese Cross Language Information Access," Program: electronic library and information systems, Emerald Group Publishing Limited, vol. 46, no. 4, p. 429 - 457, 2012.

[21] D. Dey, V. Mookerjee, and D. Liu, "Efficient techniques for online record linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 3, pp. 373–387, 2010.

[22] M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid, "TAILOR: A Record Linkage Toolbox," In Proceeding of the 18th IEEE International Conference on Data Engineering (ICDE'02), San Jose, CA, USA, pp. 17-28, Feb. 26 - March 1, 2002.

[23] Ahmed K. Elmagarmid, "Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, Jan. 2007.

[24] T. El-Shishtawy. "A Hybrid Algorithm for Matching Arabic Names," International Journal of Computational Linguistics Research, vol. 4, no. 2, pp. 87-99, 2013.

[25] C. Faloutsos and K.-I. Lin, "Fastmap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets," In Proceeding of the ACM SIGMOD International Conference on Management of Data, San Jose, vol. 24, no. 2, pp. 163–174, May 1995.

[26] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," American Statistical Society, vol. 64, no. 328, 1969.

[27] G. D. Forney Jr., "Generalized Minimum Distance Decoding", IEEE Transactions on Information Theory, vol. IT-12, pp.125 -131, 1966.

[28] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, & Haussler. "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," Bioinformatics, Oxford Journals, vol. 16, no. 10, pp. 906–914, May 2000.

[29] L.E. Gill, "OX-LINK: The Oxford Medical Record Linkage System," In Proceeding of International Record Linkage Workshop and Exposition, Arlington, VA, Washington, USA, National Acadmy Press, pp. 15-33, March 20-21, 1997.

[30] L. Gravano, et al., "Text Joins in an RDBMS for Web Data Integration," In Proceeding of the 12th International World Wide Web Conference (WWW12), Budapest, Hungary, pp. 90-101, 20-24 May, 2003

[31] W. Heeringa, P. Kleiweg, C. Gooskens and J. Nerbonne. "Evaluation of String Distance Algorithms for Dialectology," In Proceedings of the Workshop on Linguistic Distances, Association for Computational Linguistics, Sydney, Australia, pp. 51-62, July 17-21, 2006.

[32] A. Higazy, T. El Tobely, A.H Yousef and A. Sarhan, "Web-based Arabic/English Duplicate Record Detection with nested Blocking Technique" In Proceeding of the IEEE 8th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, pp. 313–318, Nov. 26-28, 2013.

[33] James W. Hunt and Thomas G. Szymanski, "A Fast Algorithm for computing Longest

Common Subsequences," Communications of the ACM, vol. 20, n.5, pp.350-353, May 1977.

[34] S. Jiampojamarn, "Grapheme-to-Phoneme Conversion and its Application to Transliteration," Doctoral Dissertation, Department of Computing Science, University of Alberta, Alta, Canada, 2011.

[35] L. Jin, C. Li, and S. Mehrotra, "Efficient Record Linkage in Large Data Sets," In Proceeding of the 8th International Conference on Database Systems for Advanced Applications (DASFAA'03), Koyto, Japan, pp. 137–146, March 26-28, 2003.

[36] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An Efficient K-means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp.881 -892, 2000

[37] M. Karthighal, S. Krishna Anand, "A Survey on Removal of Duplicate Records in Database," Indian Journal of Science and Technology, vol. 6, no. 4, April 2013.

[38] H.-s. Kim, "High Performance Record Linking," Doctoral Desertation, Computer Science and Engineering, The Pennsylvania State University, 2010.

[39] V.I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Doklady Akademii Nauk SSSR, vol. 163, no. 4, pp. 845-848, 1965.

[40] A.E. Monge and C.P. Elkan, "The Field Matching Problem: Algorithms and Applications," In Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, Portland, Oregon , pp. 267-270, Aug. 2 – 4, 1996.

[41] A.E. Monge and C.P. Elkan, "An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records," In Proceeding of the 2nd ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, Newport Beach, California, USA, pp. 23-29, Aug. 1997.

[42] G. Navarro, "A Guided Tour to Approximate String Matching,"ACM Computing Surveys, vol. 33, no. 1, pp. 31-88, 2001.

[43] S.B. Needleman and C.D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," Journal of Molecular Biology, vol. 48, no. 3, pp. 443-453, Mar, 1970.

[44] H.B. Newcombe, "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," American Journal of Human Genetics, vol. 19, no. 3, pp. 335-359, May 1967.

[45] J.C. Pinheiro and D.X. Sun, "Methods for Linking and Mining Heterogeneous Databases," In Proceeding of International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, pp. 309-313, June 1998.

[46] E. Rahm, H. H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Bulletin on Data Engineering,April 2000.

[47] V. Raman and J.M. Hellerstein, "Potter's Wheel: An Interactive Data Cleaning System," In Proceeding of 27th International Conference on Very Large Databases (VLDB '01), Roma, Italy, pp. 381-390, Sept. 11-14, 2001.

[48] E.S. Ristad and P.N. Yianilos, "Learning String Edit Distance," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522-532, May, 1998.

[49] S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," In Proceeding of the 8thACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), Edmonton, AB, Canada, pp. 269-278, July 23 - 25, 2002.

[50] We. Shaalan, K. and H. Raza, "Person Name Entity Recognition for Arabic," In Proceedings of

Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Association for Computational Linguistics: Prague, Czech Republic, pp. 17-24, JUne 23-30, 2007.

[51] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences," Journal of Molecular Biology, vol. 147, pp. 195- 197, 1981.

[52] H. Srinivasan, "Machine Learning for Person Identification with Applications in Forensic Document Analysis," Doctoral Disertation, State University of New York at Buffalo, 2008.

[53] R.L. Taft, "Name Search Techniques," Technical Report Special Report No. 1, New York State Identification and Intelligence System, Albany, N.Y., Feb, 1970.

[54] Y.R. Wang and S.E. Madnick, "The Inter-Database Instance Identification Problem in Integrating Autonomous Systems," In Proceeding of 5th IEEE International Conference on Data Engineering (ICDE '89), Los Angeles, USA, pp. 46-55, Feb. 6-10, 1989.

[55] W. E. Winkler, "Overview of Record Linkage and current Research Directions," US Bureau of the Census, Tech. Rep. RR2006/02, 2006.

[56] I. H. Witten, A. Moffat, and T. C. Bell, "Managing Gigabytes," 2nd ed. Morgan Kaufmann, 1999.

[57] S. Wu and U. Manber, "Fast Text Searching Allowing Errors," Communication of the ACM, vol. 35, no. 10, pp. 83-91, Oct. 1992.

[58] W.E. Yancey. "Bigmatch: a Program for Extracting Probable Matches from a Large File for Record Linkage," US Bureau of the Census, Tech. Rep. RRC2002/01, 2002.

[59] Ahmed H. Yousef, "Cross-Language Personal Name Mapping," International Journal of Computational Linguistics Research, vol. 4, issue 4, Dec. 2013.

[60] Ahmed H. Yousef, "Cross Language Duplicate Record Detection in Big Data," Springer International Publishing Switzerland, vol. 9, pp 147-171, 2015.

[61] Zhu, J. Lafferty and Z. Ghahramani. "Semi-supervised Learning: From Gaussian Fields to Gaussian Processes," Technical Report CMU-CS-03-175, Carnegie Mellon University, Pittsburgh, 2003.