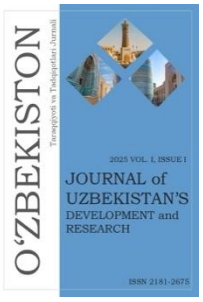


Journal of Uzbekistan's Development and Research (JUDR)

Journal home page: https://ijournal.uz/index.php/judr

FELLEGI-SUNTER METHOD FOR CONNECTING RECORDS

Ishniyazov Odil[1]

Shokirov Shodmon[2]

Alimov Xayriddin[3]

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

KEYWORDS

Fellegi-Sunter Method, Record Linkage, Block Diagram, Automated matching, Auxiliary Approach, Probabilistic Record Linkage, Matching Criteria, Key Terms, Data Processing, Data Matching Algorithm, Record Linkage Methodology, Matching Report, Data Integration, Probabilistic Model.

ABSTRACT

This scientific article explores the block diagram and key indicators of the Fellegi-Sunter method. The method provides an auxiliary and reliable approach for automated record linkage. The block diagram illustrates the detailed steps of the automated matching process, emphasizing the utilization of the Fellegi-Sunter algorithm.

2181-2675/© 2025 in XALQARO TADQIQOT LLC.
DOI: **10.5281/zenodo.15608970**

The Fellegi-Sunter method, also known as the probabilistic record linkage method, is a statistical technique used to link records from different data sources to identify and combine records that match the same entity, such as a person or business. This method was introduced in 1969 by Ivan S. Fellegi and Alan B. Sunter.

The Fellegi-Sunter method is a statistical tool used to collect data and separate it into defining elements. This method is widely used in data reporting, data extraction, data validation, and data collection. The process of linking records is important in a variety of fields, including epidemiology, demography, and the social sciences.

This method addresses the challenge of identifying and merging records corresponding to the same entity across datasets. The key components include the comparison of records using attributes, assignment of weights to attributes based on

[1] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
[2] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
[3] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

discriminatory power, a probabilistic model to calculate match probabilities, and a decision rule for classification. The approach is flexible, accounting for uncertainties in real-world data, and involves selecting thresholds to balance false positives and false negatives. Additional considerations, such as blocking, the Expectation-Maximization algorithm, and multiple comparisons, contribute to the method's robustness. Overall, the Fellegi-Sunter method plays a crucial role in diverse fields, offering a powerful tool for accurate record linkage amid imperfect and varied datasets.

In doing so, researchers often need to combine or link records from different data sets to conduct comprehensive analyses. We use this method to connect bibliographic information in information-library systems. Of course, we will analyze the association of a person, geographical place name, institution name and other data with a field element in the bibliographic database. If a person is linked by name, literature and information attributed to that author (even if created under a pseudonym) will be provided.

The main purpose of the Fellegi-Sunter method is to check, identify and study the correspondence between several data sets. This method uses the following data for statistical analysis:

• Error rate: Rate of correct data separation by mistake

• Match rate: The percentage of the number of matching data compared to the total number.

This data is a key part of the Fellegi-Sunter method, which helps calculate error and match statistics to identify suitable data.

Here's an overview of the key components of the Fellegi-Sunter method:

Comparison Vector:

Each pair of records is compared using a set of attributes (fields) that are relevant to the linkage process. These attributes might include names, addresses, birth dates, and other identifying information.

Comparison Vector Scores:

For each attribute, a comparison score is assigned based on the level of agreement or disagreement between the values in the compared records. Common scoring methods include exact match, partial match, or no match.

Weights:

Weights are assigned to each attribute based on its discriminatory power. Attributes that are more informative and less prone to errors typically receive higher weights. The weights are often determined through empirical analysis or expert judgment.

Probabilistic Model:

The method uses a probabilistic model to calculate the likelihood that a pair of records represents a match or a non-match. The model considers the comparison scores, weights, and the observed distribution of scores in the data.

Decision Rule:

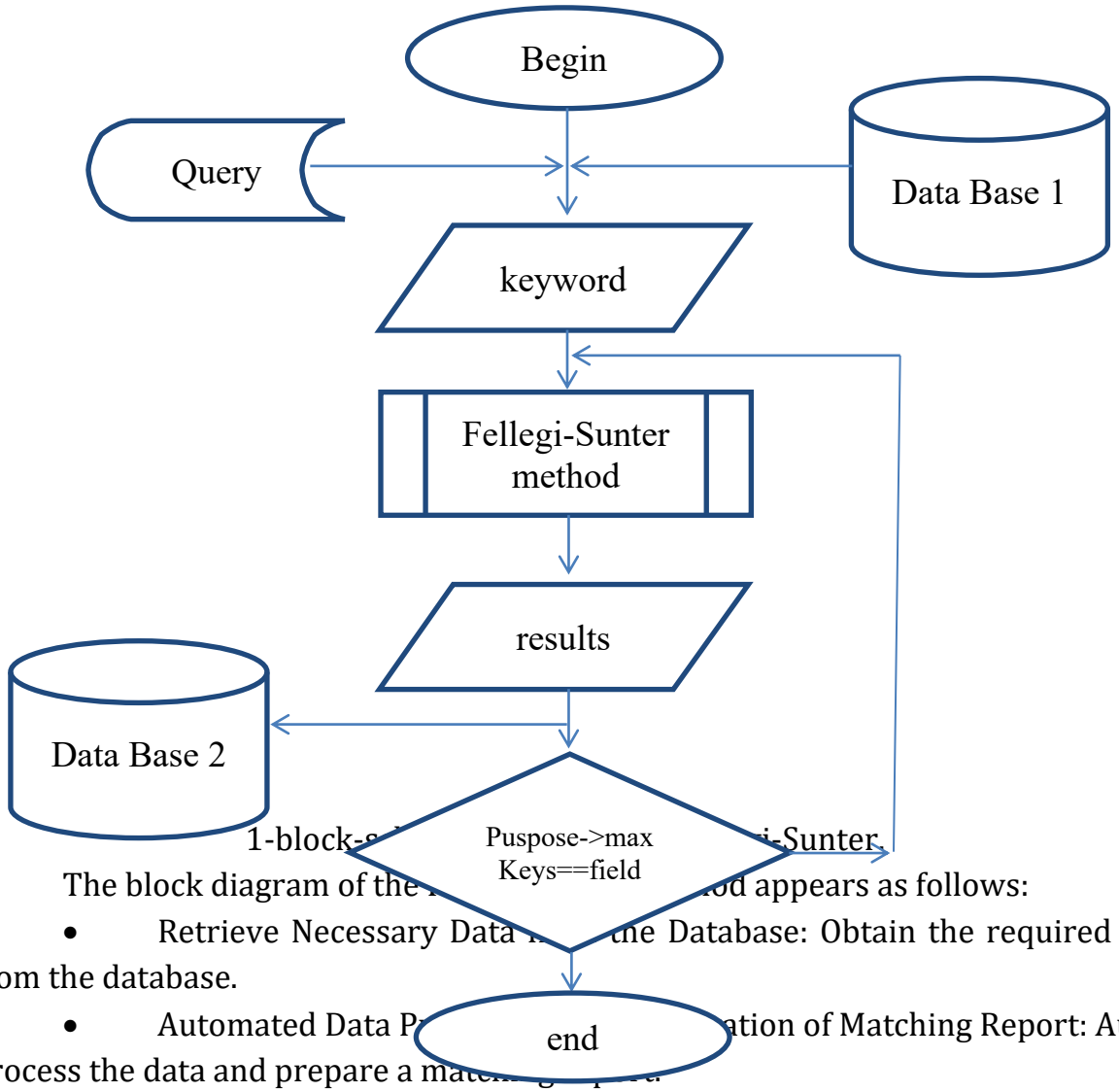
A decision rule is applied to classify record pairs as matches or non-matches based on the calculated probabilities. This rule involves setting a threshold or cutoff value, above which

pairs are considered matches and below which they are considered non-matches.

Classification Results:

The records are then classified into matches and non-matches based on the decision rule. Matches are merged, and the linked dataset is created.

The Fellegi-Sunter method allows for a more flexible and probabilistic approach to record linkage, acknowledging that perfect matches are rare in real-world datasets. It is widely used in various applications, including public health research, census data integration, and social science studies, where accurate linkage of records is essential for meaningful analysis while accounting for uncertainties and errors in the data.



1-block-s i-Sunter

The block diagram of the od appears as follows:

- Retrieve Necessary Data the Database: Obtain the required information from the database.
- Automated Data P ation of Matching Report: Automatically process the data and prepare a ma

- Input Key Terms and Matching Criteria: Input key terms and matching criteria required for the matching process.
- Implement the Fellegi-Sunter Algorithm: Execute the Fellegi-Sunter algorithm, which manages the matching process based on key terms and criteria.
- Update Matching Report and Prepare Results Report: Update the matching report with the algorithm's results and prepare an overall results report.
- Input Matching Report Responses into the Database: Input the responses from the matching report into the database.
- Utilize Responses for the Next Iteration of the Fellegi-Sunter Algorithm: Use the responses to inform the next iteration of the Fellegi-Sunter algorithm. This step is often automated.
- Verify the Purpose Suitability of Matching Report Responses Before Re-execution: Prior to re-execution, verify the suitability of matching report responses for the intended purpose.
- Re-execute and Update Matching Report: Re-execute the Fellegi-Sunter algorithm with any necessary adjustments and update the matching report.
- For Highly Automated Processes, Input New Data into the Database Automatically: For processes heavily reliant on automation, new data is automatically input into the database.

This block diagram illustrates a supportive and indicative method for record linkage within the structure of a database. It is particularly useful for automating and streamlining the matching process through a database-driven approach.

List of references:

1. Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183–1210.
2. Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic Linkage of Vital Records. Science, 130(3381), 954–959.
3. Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, 354–359.
4. Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.
5. Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data Quality and Record Linkage Techniques. Springer.