The Peerian Journal

Open Access | Peer Reviewed

Volume 38, January, 2025

ISSN (E): 2788-0303

Website: www.peerianjournal.com

Email: editor@peerianjournal.com

Data segmentation methods and algorithms

Nurmamatov Mekhriddin Kahramonovich

PhD, Associate Professor, Samarkand State University

mehriddinnur@gmail.com

Abstract: With the rapid growth of information volume, almost all databases have data redundancy or insufficiency. In the process of data matching, quality issues are particularly important. In this case, the combination of data properties such as accuracy, completeness, relevance, viability, availability, and reliability determines the quality indicators in the database. As a result of the study of segmentation methods and algorithms, it can be used to match or remove duplicate attribute values. One of the main parts of segmentation is the implementation of names and addresses using rule-based methods. Implementing segmentation based on this technology significantly increases productivity.

Key words: Segmentation, Attribute, Matching, Markov Models, Probability, Database, Area.

Introduction

In recent years, the application and creation of new technologies that enable efficient processing, management, and analysis of large datasets has become a crucial issue, considering the ever-increasing volume of databases generated not only by enterprises and government organizations but also by individuals. Currently, numerous studies are being conducted in the fields of data storage and creation. Improving the quality of many large information systems and data searches, enriching existing data sources, or implementing processing in a unified database ensures the achievement of expected results in a time and cost-effective manner [1]. This process requires the integration and adaptation of data from multiple sources. One of the key stages in the adaptation process is segmentation.

Data quality issues related to data customization.

Currently, there are cases of insufficient or redundant (repetitive) data in almost all databases. Due to a lack of data, various companies and government organizations miss out on substantial revenues. The principle of garbage-in-garbage-out applies to any type of data processing, analysis, and decision-making. This means that low-quality input data generates equivalent or even lower-quality output data [2]. To customize the records in the database, it should include the following properties:

Accuracy. Do the attribute values meet the specified requirements to eliminate duplication in data customization? Were the rules followed when recording or entering data? Has the data input undergone verification by the system administrator (responsible person)?

Completeness. How complete is the data? How many missing attribute values (empty fields) are there in the database? Are the reasons for missing attributes known? Are there sufficient attributes for data customization?

The Peerian Journal

Open Access | Peer Reviewed

Volume 38, January, 2025

ISSN (E): 2788-0303

Website: www.peerianjournal.com

Email: editor@peerianjournal.com

Consistency. How consistent are the database records (one or more) in the data matching process? The format of identified attributes in databases may change over time. Are there duplicates for the same person in the matching process (if the person changes their place of residence, obtains a new citizenship registration certificate, etc.)?

Timeliness. How long ago was the data created? Were the records in the database being matched recorded at the same time or not? This process can be a crucial step for matching. Typically, personal information such as people's phone numbers, surnames, and addresses change over time. In such cases, if the records being matched were recorded at different times, this should be taken into account during the data matching process.

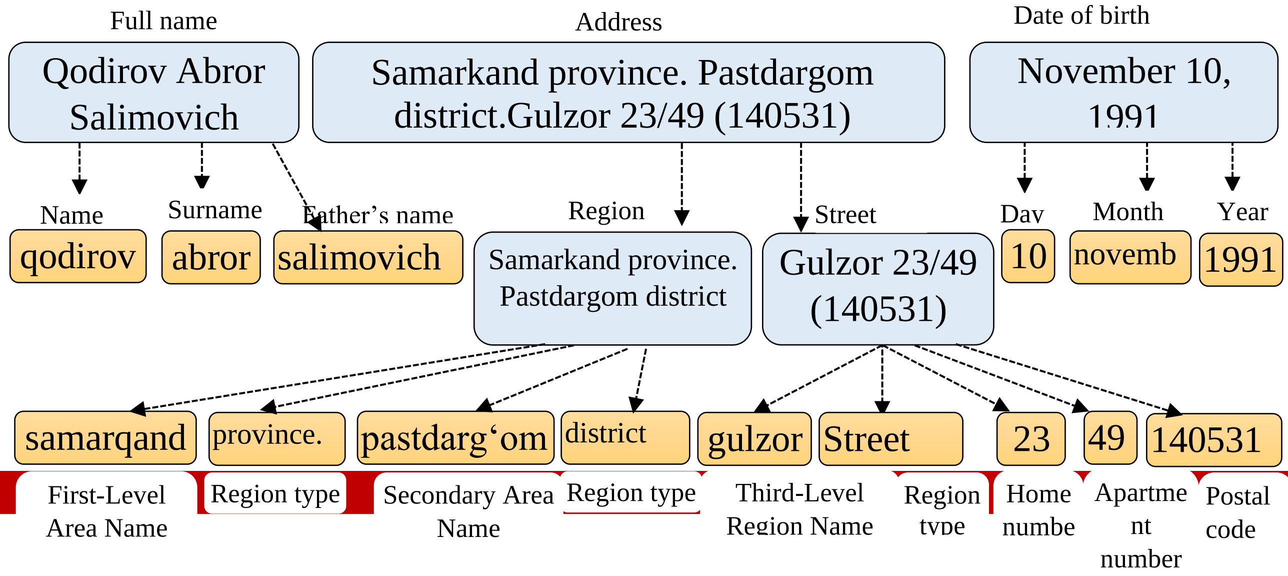
Availability. Are the records being matched present in the databases? Are there attributes in the databases that encompass all the properties necessary for complete comparisons? For example, if the databases only contain names and surnames, but lack addresses and other records, comparison work cannot be carried out.

Reliability. Can the records in the database being studied be considered accurate and authentic?

Accuracy and consistency are crucial measures for data collection, identifying discrepancies, and resolving them. The main part of the adaptation process also involves stages of indexing, comparison, and classification. When the records in the databases undergoing adaptation are of excellent quality, direct merging can be carried out without the need for indexing or comparison techniques. However, in an unorganized database, the aforementioned methods and rules are used [3].

Table 1. Preliminary database.

ID	Full name	Address	Date of Birth
a1	Qodirov Abror Salimovich	Samarkand province. Pastdargom district. Gulzor 23/49 (140531)	November 10, 1991



Based on the technologies mentioned above, the input characters in the entries of Table 1 were removed and all letters were converted to lowercase. During the standardization process, abbreviations were transformed into their corresponding standard forms. As a result, the process of segmenting data into output fields and subsequent deduplication or data matching was carried out (Table 2).

Table 2. Standardized database.

ID	Last name	First name	Father's name	First-Level Area Name	Region type	Second-Level Area Name	Region type	Day	Month	Year	Third-Level Area Name	Region type	Me nomber	Apartment number	Postal code
a1	qodirov	abror	salimovich	samarqand	province.	pastdarg'om	district	10	november	1991	gulzor	street	23	49	140531

Rule-based segmentation approaches

In the field of data personalization, rule-based methods for segmenting names and addresses have been improving for several decades. The basic idea of these technologies is to process a set of labels from left to right or right to left.

A rule-based approach works best for areas that are made up primarily of a small number of characters, such as phone numbers, house numbers, street numbers, or postal codes. Developing effective and precise rules-based technologies for addresses is quite challenging, as the sequence of characters representing addresses is quite large. This requires the creation of a complex and large set of rules.[5]

The rule-based system consists of two parts. The first is a set of rules in the form of "If the condition is met, the process begins".[5] In this case, the matching of a certain sign or sequence of signs is checked, and it involves outputting the matching signs to new fields.

A rule-based system is typically implemented by a system developed by programmers, using standardized data. Records that do not match any rule are then used to develop additional rules. Eliminating such records requires the implementation of input standards. Various programming methods exist for expressing rules, which can be implemented using scripts in programming languages such as SQL, Java, or C++. Rules can be learned automatically if training data in the form of correctly segmented input examples is available. Such training data may be pre-prepared or the data may be segmented. In general, a rule learning system incorporates a set of several partial rules. Below is a partial code of the rule-based system.

```
for i in range(len(t)):
   if t[i] == 'TI':
      title = o[i]
   elif t[i] == 'PR':
      ism_add = o[i]
   elif t[i] == 'GM' and i + 1 < len(t) and t[i + 1] == 'SN':
      name = o[i]
      fam = o[i + 1]
   elif t[i] == 'GF' and i + 1 < len(t) and t[i + 1] == 'SN':
      name = o[i]
      fam = o[i + 1]
   elif t[i] == 'SN' and i + 1 < len(t) and t[i + 1] == 'GM':
      name = o[i + 1]
      fam = o[i]
   elif t[i] == 'SN' and i + 1 < len(t) and t[i + 1] == 'GF':
      name = o[i + 1]
      fam = o[i]
   elif t[i] == 'UN' and i + 1 == L and t[i + 1] == 'SN':
      name = o[i]
      fam = o[i + 1]
   elif t[i] == 'SN' and i + 1 == L and t[i + 1] == 'UN':
      name = o[i + 1]
      fam = o[i]
```

From the above code, it can be seen that the sequence of abbreviations is denoted by o[i], and the corresponding sequence of tags is denoted by $t[i]$. $1 \le i \le L$ and L represent the number of abbreviations in the name input value.

We are given a dataset D consisting of n records for processing, and $d_1, d_2, ..., d_n$ is a sequence of symbols formed from the output fields. The primary objective is to learn k $r_1, r_2, ..., r_k$ rules that cover all the entries in the D set. Each r rule must cover all records in the D set in the form of $s(r)$ subsets, which is also known as rule coverage. Each rule is performed separately for a subset because $s(r)$ is not applicable to every entry. A small set of correctly learned records is denoted by $s'(r) \subseteq s(r)$. The accuracy of the rule is calculated as $p = |s'(r)| / |s(r)|$ [5].

The primary objective of studying a rule-based system is to create a set of rules with good coverage and high accuracy, which can then be used for segmenting unsegmented input records. Finding an optimal set of rules for a given training dataset is challenging; therefore, practical rule-learning algorithms are based on heuristic approaches.

Heuristic algorithms are based on two approaches:

- ✓ The bottom-up approach is reliable and initially covers only a single training record. It can be generalized by removing parts of conditions, but this may lead to decreased accuracy.
- ✓ The top-down approach is considered to have low accuracy as it covers numerous training records. It is refined by adding additional tests until the desired result is achieved.

To implement these approaches, methods and algorithms such as Rapier, (LP) 2, FOIL, and WHISK have been developed [5,6].

Segmentation based on the Hidden Markov model

Typically, in hidden Markov models, the beginning and end are considered special states. Starting from the initial state, the output symbol sequence $O = o_1, o_2, \ldots, o_k$, trained in hidden Markov models, undergoes $k-1$ transition states until reaching the final stage. The o_i value in the i th state $1 \le i \le k$ is based on the probability distribution of output symbols in that state. The start and end states do not store hidden Markov models, as no output symbols are generated in these states. Instead of a single initial state, a list of initial probability states is used, which provides the likelihood of the sequence starting from a specific state [7,8].

In hidden Markov algorithms, various paths can produce the same output value. However, the transition probabilities between stages remain diverse. Considering a specific sequence of output values, these hidden Markov models determine probabilistic path coefficients for segmentation. In this context, using the dynamic programming approach, the Viterbi algorithm is considered an efficient method for calculating the probabilistic path for a given sequence of output symbols [9,10].

Practical part

Table 3.

Transition extrema in a hidden Markov model

	Within the region						
	Street number	Street name	Street type	Neighborhood name	Region name	Postal code	End
Start	0.85	0.1	0.0	0.05	0.0	0.0	0.0
Street number	0.03	0.97	0.0	0.0	0.0	0.0	0.0
Street name	0.0	0.03	0.9	0.07	0.0	0.0	0.0
Street type	0.0	0.0	0.0	0.93	0.04	0.03	0.0
Neighborhood name	0.0	0.0	0.0	0.03	0.35	0.45	0.17
Region name	0.0	0.0	0.0	0.0	0.0	0.85	0.15

Postal code	0.0	0.0	0.0	0.0	0.1	0.0	0.9

Data adaptation is repeated across all training records in the learning process and adjusts the transition and output probabilities according to their output field and tag sequence. For example, the probability of transitioning from the "street type" state to the "neighborhood name" state in Table 3 is 0.93, which means that 93% of all training records contain this sequence of two output fields[11].

Conclusion

The study investigated methods and models for segmenting record fields in large databases, and the determination of quality indicators of such properties as accuracy, completeness, relevance, viability, availability, and reliability of data was presented. As a result of the study of segmentation methods and algorithms, the possibility of using them to match or remove duplicates of labeled attribute values was explained with practical problems. The solution of the practical problem was proven by calculating the transition and exit probability coefficients between the sets of states of hidden Markov models.

Literature

1. Akhatov A., Nurmamatov M., Nazarov F. “Intelligent modeling and optimization of processes in the labour market” *Artificial Intelligence, Blockchain, Computing and Security - Proceedings of the International Conference on Artificial Intelligence, Blockchain, Computing and Security, ICABCS 2023*, 2024, 2, страницы 694–699.

2. Batini, C., Scannapieco, M.: Data quality: Concepts, methodologies and techniques. Data-Centric Systems and Applications. Springer (2006).

3. Nurmamatov M.Q., Sariyev Sh.N., Genetik algoritmlar asosida turli sinfli ma’lumotlarni o‘zaro moslashtirish algoritmlari. Sh.Rashidov nomidagi Samarqand Davlat Universiteti Ilmiy axborotnomasi. 3-son (145/1) aniq va tabiy fanlar yo‘nalishi. 77-83 b.

4. Axatov A.R., Nurmamatov M.Q., Nazarov F.M. 2022. “Mathematical Models of Coordination of Population Employment in the Labor Market” // Ra journal of applied research. India / –Vol. 8, Issue 2. – Pp. 111–119. doi:https://doi.org/10.47191/rajar/v8i2.09

5. Sarawagi, S.: Information extraction. Foundations and Trends in Databases 1(3), 261–377. (2008)

6. Prasad, K., Faruquie, T., Joshi, S., Chaturvedi, S., Subramaniam, L., Mohania, M.: Data cleansing techniques for large enterprise datasets. In: SRII Global Conference, pp. 135–144. San Jose, USA (2009)

7. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)

8. Christen, P.: Probabilistic data generation for deduplication and data linkage. In: IDEAL, Springer LNCS, vol. 3578, pp. 109–116. Brisbane (2005)

9. Churches, T., Christen, P., Lim, K., Zhu, J.X.: Preparation of name and address data for record linkage using hidden Markov models. BioMed Central Medical Informatics and Decision Making 2(9) (2002)

10. Akhatov A.R., Nurmamatov M.Q., Mardonov D. 2020. "Mathematical models of the process of monitoring the social status and employment of the population", Scientific and technical journal of the Fergana Polytechnic Institute. -Volume 24, No. 5. -pp. 150–157.
11.Nurmamatov, M., Kulmirzayeva, Z. "Development of an Intelligent System for Optimization of Employment Information Using Genetic Algorithms" *AIP Conference Proceedings*, 2024, 3147(1), 040006. https://doi.org/10.1063/5.0210279